UNITED STATES DEPARTMENT OF COMMERCE
Bureau of the Census
Washington, DC 20233
MASTER FILE

MAY 05 1999

DSSD CENSUS 2000 PROCEDURES AND OPERATIONS MEMORANDUM SERIES R-1

MEMORANDUM FOR Howard Hogan
Chief, Decennial Statistical Studies Division

From: Donna Kostanich DF for DK
Assistant Division Chief, Sampling and Estimation
Decennial Statistical Studies Division

Prepared by: Ryan Cromar RC
Sample Design Team

Subject: Accuracy and Coverage Evaluation Survey: American Indian Reservations Sample Design

I. INTRODUCTION

The purpose of this memorandum is to present the sample design for American Indian Reservations (AIR) in the Accuracy and Coverage Evaluation (ACE). The planned sample allocation is designed to maximize reliability of the AIR ACE estimates while also controlling AIR weight variation among the states. Table 1 in the attachment gives the AIR block cluster sample allocation for each state. These allocation targets could change due to operational resource constraints or random sample size variation.

II. SAMPLE DESIGN

The following are features of the planned AIR sample design for the ACE:

- A total of 355 block clusters are allocated to AIR. Based on reference 1, we originally allocated 350 block clusters to AIR, but adjusted that to 355 to control weight variation.

- The originally planned number of 350 block clusters for AIR were allocated to each state proportional to the population of American Indians on reservations. We assumed that the distribution of American Indians on reservations across states does not change between 1990 and 2000. Since we cannot sample partial block clusters, we used standard rounding procedures to determine the number of block clusters sampled for each state. Table 2 in the attachment contains the unrounded expected number of AIR block clusters and the rounded number of AIR block clusters for each state for the originally planned 350 block clusters.

- The 355 block cluster allocation was reached by adding block clusters to Idaho, Michigan, and Oklahoma to control weight variation. The AIR weights for these states would be unacceptably high in comparison to the other states without the additional block clusters. Adding these block clusters and deducting a block cluster from Arizona give AIR a total of 355 sampled block clusters. Table 2 in the attachment presents details about this adjustment.

- Ten of the 36 states that have at least one AIR are not a part of the 355 block cluster allocation due to a small population of American Indians on reservations relative to other states. The AIR in these ten states will be sampled in the general population.

- For the 26 states which have an AIR allocation, the AIR sampling stratum will consist of both medium and large AIR block clusters. We will not do large-block subsampling in AIR. Note that medium block clusters have three to 79 housing units and large block clusters have 80 or more housing units.

- Small AIR blocks will be sampled in the state's small block stratum. However, small blocks in AIR will not be included in the general small-block subsampling operation. Also, small blocks in Tribal Designated Statistical Areas, Tribal Jurisdiction Statistical Areas, and Alaskan Native Village Statistical Areas but not in AIR will not be part of the general small-block subsampling operation to control weight variation for the American Indian poststrata, which will include all American Indians on reservations and on these other American Indian areas. Note that small block clusters have zero to two housing units.

- Tribal Designated Statistical Areas, Tribal Jurisdiction Statistical Areas, and Alaskan Native Village Statistical Areas are not a part of the 355 block cluster allocation due to low American Indian population density in these areas. We propose using native statistical area when sorting the block clusters in the nonAIR sampling strata to control sample size variation in those areas.

- A separate AIR stratum will be formed within each state consisting of block clusters on AIR. The Take Every (TE) for this stratum will be calculated as:

$$TE = \frac{\text{Number of Block Clusters in AIR Stratum}}{\text{AIR Block Cluster Sample Size}}.$$

- For block clustering, we will respect all AIR boundaries except when a block crosses an AIR boundary. If that happens, we will include the whole block as part of the AIR universe.

The planned AIR block cluster allocation that takes into account the above criteria is given in Table 1 in the attachment. With this allocation, the expected coefficient of variation (CV) for American Indians on reservations is approximately equal to 3.2 percent, which is based on the 1990 CV adjusted for sample size, weight variation, and a limited surrounding block search.

III. REFERENCE

[1] Schindler, E. (1998) "Allocation of the ICM Sample to the States for Census 2000," *Proceedings of the Survey Research Methods Section, American Statistical Association*, Alexandria, VA, American Statistical Association, to appear.

cc:

DSSD Census 2000 Procedures and Operations Memorandum Series Distribution List
ACE Implementation Team
Statistical Design Program Steering Committee Team Leaders
Sample Design Team

Table 1. AIR Block Cluster Allocation

State	Block Clusters	1990 American Indians on Reservations
Alabama[1]	0	149
Alaska	1	1,209
Arizona	113	142,238
Arkansas	0	0
California	11	13,602
Colorado	2	2,063
Connecticut[1]	0	79
Delaware	0	0
DC	0	0
Florida	1	1,517
Georgia[1]	0	16
Hawaii	0	0
Idaho	6	5,896
Illinois	0	0
Indiana	0	0
Iowa[1]	0	564
Kansas	1	988
Kentucky	0	0
Louisiana[1]	0	261
Maine	1	1,482
Maryland	0	0
Massachusetts[1]	0	1
Michigan	5	2,996
Minnesota	10	12,472
Mississippi	3	3,932
Missouri	0	0
Montana	24	30,424
Nebraska	3	3,521
Nevada	5	5,854
New Hampshire	0	0
New Jersey[1,2]	0	0
New Mexico	70	87,659
New York	5	6,272
North Carolina	4	5,388
North Dakota	12	15,284
Ohio	0	0
Oklahoma	8	6,088
Oregon	3	4,013
Pennsylvania	0	0
Rhode Island[1]	0	17
South Carolina[1]	0	124
South Dakota	27	33,931
Tennessee	0	0
Texas	1	688
Utah	7	8,577
Vermont	0	0
Virginia[1]	0	100
Washington	17	21,794
West Virginia	0	0
Wisconsin	10	12,483
Wyoming	5	5,676
Total	355	437,358

1 States contain AIR population, but not AIR sampling stratum. AIR people will be given a chance of selection in the general state sample.

2 New Jersey AIR had no population in 1990.

Table 2. AIR Block Cluster Allocation Adjustments

State	Expected Block Clusters	Block Clusters before Adjustment	Weights before Adjustment	Block Cluster Adjustment	Weights after Adjustment
Alabama[1]	0.12	0	NA		NA
Alaska	0.97	1	14.5667		14.5667
Arizona	113.83	114	11.5980	-1	11.7006
Arkansas	0.00	0	NA		NA
California	10.89	11	56.4697		56.4697
Colorado	1.65	2	54.6833		54.6833
Connecticut[1]	0.06	0	NA		NA
Delaware	0.00	0	NA		NA
DC	0.00	0	NA		NA
Florida	1.21	1	24.5333		24.5333
Georgia[1]	0.01	0	NA		NA
Hawaii	0.00	0	NA		NA
Idaho	4.72	5	73.9067	1	61.5889
Illinois	0.00	0	NA		NA
Indiana	0.00	0	NA		NA
Iowa[1]	0.45	0	NA		NA
Kansas	0.79	1	17.6000		17.6000
Kentucky	0.00	0	NA		NA
Louisiana[1]	0.21	0	NA		NA
Maine	1.19	1	15.8333		15.8333
Maryland	0.00	0	NA		NA
Massachusetts[1]	0.00	0	NA		NA
Michigan	2.40	2	168.7833	3	67.5133
Minnesota	9.98	10	43.2833		43.2833
Mississippi	3.15	3	9.7222		9.7222
Missouri	0.00	0	NA		NA
Montana	24.35	24	27.0236		27.0236
Nebraska	2.82	3	29.2667		29.2667
Nevada	4.68	5	13.9467		13.9467
New Hampshire	0.00	0	NA		NA
New Jersey[1,2]	0.00	0	NA		NA
New Mexico	70.15	70	16.0857		16.0857
New York	5.02	5	34.4867		34.4867
North Carolina	4.31	4	14.2083		14.2083
North Dakota	12.23	12	17.0167		17.0167
Ohio	0.00	0	NA		NA
Oklahoma	4.87	5	102.2800	3	63.9250
Oregon	3.21	3	16.6889		16.6889
Pennsylvania	0.00	0	NA		NA
Rhode Island[1]	0.01	0	NA		NA
South Carolina[1]	0.10	0	NA		NA
South Dakota	27.15	27	19.4728		19.4728
Tennessee	0.00	0	NA		NA
Texas	0.55	1	7.3000		7.3000
Utah	6.86	7	40.6619		40.6619
Vermont	0.00	0	NA		NA
Virginia[1]	0.08	0	NA		NA
Washington	17.44	17	55.5941		55.5941
West Virginia	0.00	0	NA		NA
Wisconsin	9.99	10	35.7833		35.7833
Wyoming	4.54	5	54.3533		54.3533
Total	350.00	349[3]		6	

1 States contain AIR population, but not AIR sampling stratum. AIR people will be given a chance of selection in the general state sample.

2 New Jersey AIR had no population in 1990.

3 After rounding, the total block clusters summed to 349 for AIR instead of 350.

UNITED STATES DEPARTMENT OF COMMERCE
Bureau of the Census
Washington, D.C. 20233

MASTER FILE

March 29, 1999

DSSD CENSUS 2000 PROCEDURES AND OPERATIONS MEMORANDUM SERIES R-3

MEMORANDUM FOR Dennis W. Stoudt
Assistant Division Chief, Processing Systems
Decennial Systems and Contracts Management Office

From: Donna Kostanich DK
Assistant Division Chief, Sampling and Estimation
Decennial Statistical Studies Division

Prepared by: Thomas Mule TM
Sample Design Team
Decennial Statistical Studies Division

Subject: Accuracy and Coverage Evaluation (ACE) Survey:
Block Cluster Sample Selection Specification

I. INTRODUCTION

This memorandum describes the selection of the initial block cluster sample for the ACE Survey. The plan is to select a national sample of 25,000 block clusters plus 5,000 small block clusters. This includes a separate sample of 355 block clusters for American Indians Reservations (AIR). An additional 480 block clusters will be selected for Puerto Rico. This sample will be provided to the Field Division for independent listing. The results from the independent listing will be used to select a reduced sample for ACE of approximately 300,000 housing units. Requirements and details of the ACE design are not known at this time. A separate operation will be specified in a future memorandum to reduce the number of ACE sample clusters.

This specification describes a two step sampling process. The first step is the selection of the initial block cluster sample. The second step is a subsample of the first-step cluster when the estimated listing workload is too high. This second step is a contingency plan. Since listing constraints are accounted for when calculating sampling parameters, expectations are low that this step of sampling will be needed.

Before the block clusters can be sampled for each state, the Universe File and Block Cluster Sampling Parameter File must be completed and approved. This procedure assumes that the stratification variables were assigned in the Universe File creation and the sampling parameters were calculated. These processes are specified in *"Accuracy and Coverage Evaluation (ACE)*

Survey: Universe File and Block Cluster Sampling Parameter File Specification." After the Decennial Statistical Studies Division (DSSD) reviews and approves the block cluster sample selection for each state then the sampled block clusters will be sent to the Geography (GEO) Division.

These specifications should be used to flowchart the process, to generate further discussion on requirements, to identify and finalize the record layouts of input and output files, and to write computer software to implement the methodology. During and after a testing phase, it is likely that changes to the specifications will be necessary.

The sections of this specification are ordered as follows:

- Section II states the assumptions and definitions of the sample selection process.
- Section III lists the input and output files for this process.
- Section IV lists the first-step sample selection process.
- Section V determines if second-step sampling is necessary.
- Section VI lists the second-step sample selection process.
- Section VII lists the sample selection output.
- Section VIII specifies the processes after the sample block clusters fields are updated by the GEO Division.

Any comments or questions should be directed to Thomas Mule (301) 457-8322 or James Farber (301)457-4282.

II ASSSUMPTIONS/DEFINITIONS

A. Block clusters have been created and verified for an entire state before sampling occurs in that state.

B. Stratification is complete. Sampling strata are specified in *"Accuracy and Coverage Evaluation (ACE) Survey: Universe File and Block Cluster Sampling Parameter File Specification."*

C. A target of 30,000 block clusters will be selected for the 50 states and the District of Columbia: 25,000 medium and large and 5,000 small. The allocation of the block clusters to the states is listed in Appendix I.

D. The reduction of ACE sample clusters will be specified in a future memorandum.

E. All numbers have been rounded to the sixth place after the decimal, xxxx.xxxxxx.

F. Small block clusters have 0-2 housing units (HUs). Medium block clusters have between 3-79 HUs. Large block clusters have 80 HUs or more.

G. For each state, there are four possible sampling strata:

1) Small Block Clusters,
2) Medium Block Clusters,
3) Large Block Clusters,
4) American Indian Reservation (AIR) Block Clusters.

Note: The fourth stratum does not exist in all states.

H. Within each sampling stratum, the block clusters will be implicitly stratified. The implicit stratification for each state was determined by research done by the Sample Design Branch staff and applied during the stratification process.

I. The Estimated Listing workload resulting from selection for a state must be ten percent greater than the budgeted listing to require the second-step sampling.

III. FILES

The following files will be used in this process.

A. INPUT FILES

1. Block Cluster Sampling Parameter File: The first-step sampling parameters, e.g. Take-Every's and Random Starts, for each state will come from this file. There is one record for each sampling stratum in a state. Appendix A has a copy of the Block Cluster Sampling Parameter File layout that was originally specified in *"Accuracy and Coverage Evaluation (ACE) Survey: Universe File and Block Cluster Sampling Parameter File Specification."*

2. Universe File: Each state's Universe File was created during the universe creation process. There is a record for each block in every cluster. Necessary records and variables for sampling will be taken from this file. Appendix B has a copy of the Universe File layout that was originally specified in the *"Accuracy and Coverage Evaluation (ACE) Survey: Universe File and Block Cluster Sampling Parameter File Specification."*

3. Sample Size Input File: For each state, the budgeted number of housing unit listings has been determined. After sampling the estimated number of listings will be determined for each state. The estimated number will be compared to the budgeted number to make sure it is within our cost constraints. The DSSD will provide to the Decennial Systems and Contracts Management Office (DSCMO) the file, 2000_TB.FIN, which contains the budgeted counts. This is a state level file. These budgeted numbers for each are documented in Appendix I. The layout for this file is in Appendix H.

4. Sample Summary File: The Sample Summary File has state-level summary statistics to track the sampling process. This file was created during the Universe File creation process. Estimates will be added to this file for each state. Appendix F has a copy of the file layout that was originally specified in *"Accuracy and Coverage Evaluation (ACE) Survey: Documentation for the Sample Summary File and Sample Design File."*

B. OUTPUT FILES

1. Block Cluster Sampling Parameter File: After sampling, the estimated number of housing units for listing will be calculated for each sampling stratum. This count will be appended to the Block Cluster Sampling Parameter File.

2. Sample Design File: The Sample Design File tracks the path that each sampled block cluster travels during the ACE sampling procedures. It is created after the first-step block cluster sample is selected and contains one record for each first-step sampled block cluster. The layout of the Sample Design File was originally specified in *"Accuracy and Coverage Evaluation (ACE) Survey: Documentation for the Sample Summary File and Sample Design File"* and is given again in Appendix C.

3. Universe File: After sampling, the Universe File will be updated to indicate if a block cluster was selected in the first or second step. The Index Numbers will be attached to the file so the sampling sort can be replicated.

   Note: For these specifications, if a block cluster is assigned a value on the Universe File, then all blocks in the cluster will receive the assigned value.

4. ACE Stratification Summary File: This file will contain summary sample estimates for each of the demographic/tenure groups in the sampling strata. This will provide summary information that can be used in determining the parameters for reducing the ACE sample cluster in the future. The layout for this file is in Appendix D.

5. ACE Sample Cluster File: This file will have a record for each block from a block cluster that remains in sample after the second step. This file will be sent to the GEO. The GEO will use this file to identify ACE sample blocks on the collection Geographic Reference File (GRF) and to create the ACE GRF. The layout for file that will be sent from the DSCMO to the GEO is in Appendix E.

6. Field Prioritization File: This file will have a record for each block cluster that remains in sample after the second step. This file will be sent to Field Division so

they can prioritize their Housing Unit Follow-up (HUFU) and Person Follow-up (PFU) workloads. The layout for this file is in Appendix G.

7. Sample Summary File: The Sample Summary File has state-level summary statistics to track the sampling process. This file was created during the Universe File creation process. A second-step sampling indicator, the number of clusters in sample after each step and housing units to list estimates will be added to this file for each state. The layout for this file is in Appendix F.

IV. FIRST-STEP SAMPLE SELECTION PROCESS

A. OVERVIEW

The first-step sample selection process will select the initial ACE block cluster sample for each state. This process will obtain the sampling parameters for each sampling stratum from the Block Cluster Sampling Parameter File. An index number will be attached to each block cluster to identify the block clusters selected during the systematic sampling. An unique ID will be assigned to the sampled block clusters and the Sample Design File will be created. After sampling, the estimated listing workloads will be generated for each sampling stratum and recorded on the Block Cluster Sampling Parameter File. These workloads will be compared to the budgeted amount to determine if the second-step of sampling is needed.

Each state, the District of Columbia and Puerto Rico will be sampled separately. The process described below applies to each state's sample selection.

B. SELECTION PROCESS

1. Sort the Universe File

The Universe File will be sorted prior to sample selection. This sorting will reduce the variability of sample size among demographic/tenure groups and ensure a fair representation of block clusters across the counties in the state.

a. For each state, use the Universe File.

b. Sort the block clusters in the following order:

- Sampling Stratum (SS)
- American Indian Country Indicator (AICIND)
- Demographic/Tenure Group (DTCODE)
- 1990 Estimated Urbanization (ECLUSURB)
- County (COUNTY)
- Geographic Block Cluster Number (GCLUST)

2. Attach First-Step Index Number to Each Block Cluster

Number the block clusters consecutively from 1 to N within each sampling stratum where N is the number of block clusters in the stratum. The assigned number is referred to as the First-Step Index Number of the block cluster. Assign the First-Step Index Number (INDEX1) of each block cluster to the Universe File.

3. Select Sample

For each of the sampling strata, select a separate systematic sample of block clusters as follows:

a. Generate a sequence of numbers $L_1,\ldots,L_n$ as follows:

- From the Block Cluster Sampling Parameter File, obtain the Random Start for Initial Block Cluster Sampling (RS1) and the Take -Every for Initial Block Cluster Sampling (TE1)

- Let $L_1 = RS1$

- Calculate $L_j = L_{j-1} + TE1$, for $j = 2$ to n where n is the largest integer such that $[RS + (n - 1) \times TE1] <= N$

- Round each L_j up to the nearest integer (an integer round to itself).

- For each block cluster in the sampling stratum:

  If the first-step index number is equal to the rounded values of L_j, $j = 1,\ldots,n$ then do the following:

  - Assign the First-Step Sample Indicator (BC1) on the Universe File equal to '1'. The block cluster was selected in sample.

  - Assign the Current Sample Indicator (CSI) on the Universe File equal to '1'. The block cluster is currently in sample.

  Otherwise, do the following:

  - Assign the First-Step Sample Indicator (BC1) on the Universe File equal to '0'. The block cluster was not selected in sample.

  - Assign the Current Sample Indicator (CSI) on the Universe File equal to '0'. The block cluster was not selected so it is not currently in sample.

  For example: if $N = 100$, $RS1 = 2.4$ and $TE1 = 7.2$, then $n = 14$. Set $L_1 = 2.4$. The generated L_js would be the sequence: 2.4, 9.6, 16.8, 24.0,...,96.0. Therefore the block clusters with First-Step Index Numbers 3, 10, 17, 24, 32,..., and 96 would be selected for the sample.

b. Compute a Check

For each sampling stratum, check the number of sampled block clusters, given by n, by calculating c, which is a check of the sampling procedures:

$$c = \left| \frac{N}{TE1} - n \right|$$

If the sampling is implemented correctly, c will be less than 1. For values of c that are not less than one and have not been resolved, contact the DSSD for review of the sampling operations.

4. Number the Selected Sample Block Clusters

In each state, sort the selected block clusters by county and geography block cluster number. Number the block clusters selected for the first-step sample using the following algorithm.

The ACE block cluster sample number (CLUST) will be a five digit number. The first digit within the five-digit cluster number will represent the Census division. There are nine Census divisions. The remaining four digits in the five-digit cluster number will be a sequence number. Appendix J contains the range of cluster numbers allocated to each state within a division.

For each state, start with the lowest value in its allocated range of cluster numbers. Assign this to the first cluster in the sort. Increment the cluster number by one and assign it to the next cluster and so on. Do not assign cluster numbers to any block cluster that was not selected in the first step of sampling.

For example, Texas has a range of ACE cluster numbers between 54001 and 57999. Assign 54001 to the first cluster, 54002 to the second cluster, and so on.

Assign the check digit (DIGIT) for the ACE block cluster sample number. The DSCMO will use the Double-Add-Double Check-digit Algorithm to assign the check digit. Appendix K documents this algorithm.

5. Create the Sample Design File

The Sample Design File tracks the path that each sampled block cluster travels during the ACE sampling procedures. It is created after the first-step block cluster sample is selected and contains one record for each sampled block cluster.

The layout of the Sample Design File was originally specified in *"Census 2000 Accuracy and Coverage Evaluation: Documentation for the Sample Summary File and Sample Design File."* The layout is given again in Appendix C.

a. Create the file ACE_SDFV1_<mmddyy>.<SA> where <SA> is the state abbreviation (i.e AL, AK, AZ, etc.) for the state being sampled and <mmddyy> is the date of sample selection.

b. Create a record for each first-step sampled block cluster. Put the following fields on the file:

Variable Description	Name
Census Region	REGION
Census Division	DIV
State code (FIPS)	STATE
County Code (FIPS)	COUNTY
Interim Tract (Pseudo-Tract)	ITRACT
ACE Block Cluster Number	CLUST
ACE Block Cluster Check Digit	DIGIT
Geography Block Cluster Number	GCLUST
TEA Group	TEAG
Sampling Stratum	SS
Demographic/Tenure Group Code	DTCODE
Demographic/Tenure Group Label	DTLABEL
Number of Housing Units for Sample Des.	NHU
Number of 2000 MAF Housing Units	NHUM
Number of 1990 Estimated Housing Units	NHU90
First-Step Index Number	INDEX1
Estimated Urbanicity of Block Cluster	ECLUSURB
American Indian Country Indicator	AICIND
Size Category	SIZCAT
Current Sample Indicator	CSI
Initial Block Cluster Sampling Indicator	BC1
Random Start for Initial Block Cluster Sampling	RS1
Take-Every for Initial Block Cluster Sampling	TE1

c. For each record, set BC1 equal to '1' to indicate that the block cluster was sampled during the first step.

d. For each record, set CSI equal to '1' to indicate the block cluster is currently in sample.

C. CALCULATE ESTIMATED WORKLOADS FOR SAMPLING STRATA

For each sampling strata, calculate the estimated listing workload (INMHUL) by summing the number of housing units of the sampled block clusters in each strata.

$$INMHUL = \sum_{i=1}^{n} NHU_i$$

where n is the number of block clusters in sample in the stratum and NHU_i is the number of housing units in the block cluster.

Update the INMHUL for each sampling stratum on the Block Cluster Sampling Parameter File.

D. SECOND-STEP BLOCK CLUSTER SAMPLING DETERMINATION

The second-step block cluster sampling will be done when the state's estimated listing workload in the medium and large sampling strata (NHUL1_ML) is greater than 110 percent of the budgeted listing workload (BLIST). Otherwise, the second-step sampling is not needed.

Obtain the budgeted listing workload for each state from the 2000_TB.FIN file. The budgeted listing workload for each state is listed in Appendix I.

Determine the state's estimated listing workload in the medium and large sampling strata. as follows:

$$NHUL1\_ML = \sum_{i \in Medium,\, Large} INMHUL_i$$

where *i* is the sampling stratum.

The second-step block cluster sampling will be needed if the Listing Workload Ratio is greater than or equal to 0.10:

$$\frac{NHUL1\_ML - BLIST}{BLIST} \geq 0.10$$

For documentation purposes, the total state estimated listing workload after first-step sampling will be calculated. This is the sum across all of the sampling strata.

$$NHUL1 = \sum_{i=1}^{4} INMHUL_i$$

Using the Sample Summary File, assign the State Estimated HUs In Sample to list After First-Step Sampling (NHUL1) and the Estimated HUs In Sample To List In Medium And Large Strata After First-Step Sampling (NHUL1_ML) to the state record being sampled.

If second-step block cluster sampling is deemed necessary, then continue the process by going to section V.B. If second-step block cluster sampling is not needed, then go to section VI.C.

E. DETERMINE WHICH SAMPLING TO DO SECOND STEP

Since a second-step sampling is necessary in the state, the next step is to identify in which sampling strata this operation will be done. The small and AIR strata are exempt from this step of sampling. Ideally, the second-step sampling would be limited to the large sampling stratum. However, if the second-step sampling rate is too low, then this causes differential weighting and sample size concerns. In which case, the second step will be done in both the medium and large sampling strata.

In order to determine which sampling strata to do the second step, perform the following calculation and analysis:

Calculate Check1, C1.

$$C1 = \frac{INMHUM_{Large}}{BLIST - INMHUM_{Medium}}$$

C1 is the second-step TE when only doing this step of sampling in the large stratum. If the product of the first and second-step TEs in the large stratum is greater than the medium TE, then we want to do the second step in both the medium and large strata.

Calculate the critical value as the ratio of the first-step TEs for the medium and large strata.

$$C = \frac{TE1_{medium}}{TE1_{large}}$$

If C1 is less than or equal to C then do the second step in the large stratum only; otherwise, do the second step in both medium and large strata. In other words, subsampling in the large stratum only will not be done when the overall sampling rate after both the first and second steps of sampling for the large stratum is less than for medium.

VI Second-Step Block Cluster Sampling

A. OVERVIEW

The purpose of the second step is to subsample the first-step sampled block clusters if the first-step results in an unusually high amount of housing units to list.

The second step occurs only if the expected number of housing units in the medium and large strata is at least ten percent larger than the number of housing units budgeted for listing. The second-step sampling process will be similar to the first step. The first-step sampled block clusters will be sorted by the original order of selection. A second-step index number will be attached to each block cluster to identify the block clusters selected during the systematic sampling. The Sample Design File, Block Cluster Sampling Parameter File and the Universe File will be updated to reflect the second-step sampling.

For a state needing the second step, subsampling strata determined by the check in Section V.B. will be processed using the specifications in Section VI..B. Non-subsampling strata will be processed using the specifications in Section VI.C.

If the estimated number of housing units is not ten percent larger than the number of housing units budgeted for listing then all first-step sampled block clusters will remain in sample. Since no strata are being subsampled, all strata in the state will be processed using the specifications in Section VI.C.

B. SECOND-STEP SAMPLING STRATA

Based on the check in section V.A., a state with over ten percent of the budget listing workload will be subsampled. The check in section V.B., indicates which strata will have second-step sampling.

1. Calculate the Second-Step Take-Every, TE2:

   If doing the second-step sampling only in the large stratum, then

$$TE2 = \frac{INMHUM_{Large}}{BLIST - INMHUM_{Medium}}$$

If doing the second-step sampling in both the medium and large strata, then

$$TE2 = \frac{NHUL1\_ML}{BLIST}$$

for both the medium and large strata.

2. Select the Sample:

Do the second-step sampling separately in each stratum as follows:

a. On the Block Cluster Sampling Parameter File, set the Indicator for Second Step of Block Cluster Sampling (I2) equal to '1' for the sampling stratum. This indicates that the second step is needed in this stratum.

b. Sort the block clusters selected in the first step by First-Step Index Number (INDEX1).

c. Number the first-step sampled block clusters consecutively from 1 to M. (This number is referred to as the Second-Step Index Number of the block cluster)

d. Select a systematic sample of block clusters as follows:

- Generate a random number (RN2) between 0 and 1 ($0 < RN2 \leq 1$).
- Calculate the Second-Step Random Start, $RS2=RN2 \times TE2$.
- Generate a sequence of numbers $L_1,\ldots,L_m$ as follows:
- Let $L_1 = RS2$
- Calculate $L_j = L_{j-1} + TE2$, for $j = 2$ to m where m is the largest integer such that $[RS2 + (m - 1) \times TE2] <= M$
- Round each L_j up to the nearest integer (an integer round to itself).

e. Compute a Check

Check the number of selected second-step block clusters by calculating c, which is a check of the sampling procedures:

$$c = \left| \frac{M}{TE2} - m \right|$$

If the sampling is implemented correctly, c will be less than 1. For values of c that are not less than one and have not been resolved, contact the DSSD for review of the sampling operations.

If the sampling is not implemented correctly, do not proceed with the remaining steps in this part until it is resolved.

f. Each block cluster with a Second-Step Index Number equal to the rounded values of L_j, $j = 1,...,m$, are the selected second-step sample block clusters. Do the following for each second-step sampled block cluster:

i. Use the Sample Design File and find the block cluster record.

- Assign the Second-Step Block Clustering Sampling Indicator (BC2) equal to '1'. This indicates that the block cluster was selected in the second step.

- Assign the Second-Step Index Number (INDEX2).

- Assign the Random Start for Second-Step Block Cluster Sampling (RS2).

- Assign the Take-Every for Second-Step Block Cluster Sampling (TE2).

- Calculate the unbiased weight after block cluster sampling (WEIGHTBC):

$$WEIGHTBC = TE1 \times TE2$$

Assign the unbiased weight after block cluster sampling (WEIGHTBC) to the file.

ii. Use the Universe File and find the block cluster record.

- Assign the Second-Step Block Clustering Sampling Indicator (BC2) equal to '1'. This indicates that the block cluster was selected in the second step.

- Put the Second-Step Index Number (INDEX2) on the Universe File.

g. If the Second-Step Index Number of the block cluster does NOT match one of the rounded values of L_j, $j = 1,...,m$, then the block cluster is no longer in the sample. Do the following for each of these block clusters:

i. Use the State Sample Design File and find the block cluster record.

- Assign the Second-Step Block Clustering Sampling Indicator (BC2) equal to '0'. This indicates that the block cluster was NOT selected in the second step.

- Change the Current Sample Indicator (CSI) equal to '0'. The cluster is no longer in sample.

- Put the Second-Step Index Number (INDEX2) on the file.

- Put the Random Start for Second-Step Block Cluster Sampling (RS2) on the file

- Put the Take-Every for Second-Step Block Cluster Sampling (TE2) on the file

- Assign the Unbiased Weight After Block Cluster Sampling (WEIGHTBC) a value of '   ' (12 blanks). This block cluster was not selected.

ii. Use the Universe File and find the block cluster record.

- Assign the Second-Step Block Clustering Sampling Indicator (BC2) equal to '0'. This indicates that the block cluster was NOT selected in the second step.

- Change the Current Sample Indicator (CSI) equal to '0'. The cluster is no longer in sample.

- Put the Second-Step Index Number (INDEX2) on the Universe File.

h. Make the following updates for block clusters not selected in the first-step on the Universe File. These records have First-Step Block Cluster Sampling Indicator (BC1) set equal to '0'. This step is to maintain clear documentation of the Universe File for future sample selections.

   i. Assign the Second-Step Block Cluster Sampling Indicator (BC2) equal to '0'. This indicates the cluster is not in sample.

   ii. Set the Second-Step Index Number (INDEX2) equal to '   ' (five blanks). The clusters were not involved in the subsampling.

i. Count the number of second-step sampled clusters in the sampling stratum.

   Assign this count to the Clusters In Sample To List (CLUSL) field for the sampling stratum on the Block Cluster Sampling Parameter File.

j. Count the number of housing units to be listed for the second step sample clusters in the sampling stratum.

   Assign this count to the Housing Units In Sample To List (NMHUL) for the sampling stratum on the Block Cluster Sampling Parameter File.

k. Update the Random Start for Second-Step Block Cluster Sampling (RS2) field and the Second-Step Take-Every (TE2) field for the sampling stratum on the Block Cluster Sampling Parameter File.

C. NO SECOND STEP NEEDED IN STRATUM

For state where the second-step process is not required, all first-step sampled clusters in all of the strata remain in sample. These states go through the following process. For states where subsampling occurs, the strata not involved in subsampling go through the following process.

1. Make the following updates to the Block Cluster Sampling Parameter File for each of the sampling strata not needing the second-step sampling:

   a. Assign Take-Every for Second-Step Block Cluster Sampling (TE2) equal to 1.000000.

b. Assign Random Start for Second-Step Block Cluster Sampling (RS2) equal to 1.000000.

c. Assign Indicator for Second Step of Block Cluster Sampling (I2) equal to '0'. The second step was NOT needed in these sampling strata.

d. Set the Clusters in Sample to List (CLUSL) equal to the Initial Clusters in Sample to List (ICLUSL).

e. Set the Housing Units in Sample to List (NMHUL) equal to the Initial Housing Units in Sample to List (INMHUL).

2. Make the following updates for all records in each sampling strata not involved in second-step subsampling on the Sample Design File:

a. Assign the Second-Step Block Cluster Sampling Indicator (BC2) equal to '1'. This indicates that the block cluster was retained in the sample.

b. Set the Second-Step Index Number (INDEX2) equal to '   ' (5 blanks). Second-Step Index Numbers are only assigned if subsampling in the stratum is necessary.

c. Assign Take-Every for Second-Step Block Cluster Sampling (TE2) equal to 1.000000.

d. Assign Random Start for Second-Step Block Cluster Sampling (RS2) equal to 1.000000.

e. The Unbiased Weight after Block Cluster Sampling (WEIGHTBC) is equal to the First-Step Take-Every (TE1).

$$WEIGHTBC = TE1$$

Assign the Unbiased Weight after Block Cluster Sampling (WEIGHTBC) to the file.

3. Make the following updates for first-step sampled block clusters not involved in second-step subsampling on the Universe File. These records have First-Step Block Cluster Sampling Indicator (BC1) set equal to '1'.

a. Assign the Second-Step Block Cluster Sampling Indicator (BC2) equal to '1'. This indicates the cluster is still in sample.

b. Set the Second-Step Index Number (INDEX2) equal to '   ' (five blanks). Second-Step Index Numbers are only assigned if subsampling in the stratum is necessary.

4. Make the following updates for first-step non-sampled clusters in the state on the Universe File. These records have First-Step Block Cluster Sampling Indicator (BC1) set equal to '0'.

   a. Assign the Second-Step Block Cluster Sampling Indicator (BC2) equal to '0'. This indicates the cluster is not in sample.

   b. Set the Second-Step Index Number (INDEX2) equal to '   ' (five blanks). Second-Step Index Numbers are only assigned if subsampling in the stratum is necessary.

VII. OUTPUT

A. ACE STRATIFICATION SUMMARY FILE

After the listing of HUs in each cluster, cluster reduction will be done to reach the ACE sample size. The ACE Stratification Summary File will provide the information for developing this reduction. This file will have a record for each demographic/tenure group code (DTCODE) in every sampling stratum in the state. The file will provide the estimated number of sampled housing units and demographic/tenure people by summing over the sampled block clusters. The layout of the file is in Appendix D.

For each demographic/tenure group code, count the following:

1. HUs (NHU)
2. Black/Owner People
3. Black/Renter People
4. Hispanic/Owner People
5. Hispanic/Renter People
6. Asian/Owner People
7. Asian/Renter People
8. Hawaiian and Pacific Islander/Owner People
9. Hawaiian and Pacific Islander/Renter People
10. American Indian Reservation /Owner People
11. American Indian Reservation/Renter People
12. American Indian Not On Reservation/Owner People
13. American Indian Not On Reservation/Renter People
14. White and Other/Owner People
15. White and Other/Renter People

Note: Use the people population counts from the Universe File

B. ACE SAMPLE CLUSTER FILE

After sampling, the block Type of Enumeration Area (TEA), cluster TEA and Local Census Office will need to be updated. The block TEA and cluster TEA may have changed since clustering. The Local Census Office boundaries were not available when clustering began. The DSCMO will create the ACE Sample Cluster File that will be sent to the GEO. For all block clusters that remain in sample after the second step, there will be one record for each block in the cluster. The layout is in Appendix E. The GEO will use this file to create the ACE GRF.

Create the ACE Sample Cluster File with the following variables for each block:

Variable Description	Name
State	STATE
Local Census Office	LCO
County	COUNTY
Tract	ITRACT
ACE Cluster Number	CLUST
Check Digit	DIGIT
Sampling Strata	SS
Demographic/Tenure Group Code	DTCODE
2000 Collection Block	BK2K
Geography Cluster Number	GCLUST
Cluster Size Recode from Geography	GSIZE
Number of Housing Unit in Cluster	NHU
Number of Housing Units in Block	NHB
Number of 2000 MAF Housing Units in Block	NHUMB
Number of 1990 Estimated Housing Units in Block	NHU90B
Total Persons in the Cluster	NP

C. ACCESS TO FILES FOR REVIEW

The sampling process can be reviewed if access is provided to the Sample Design File, the Block Cluster Sampling Parameter File, the Universe File and the ACE Sample Cluster File from our DMBA01 Alpha machine. Notify the DSSD Sample Design staff when the files are available and where they are located.

VIII. UPDATE SAMPLE DESIGN FILE AND CREATE FIELD PRIORITIZATION FILE

The DSCMO will use the ACE GRFs to update the Sample Design File. After the file is received from the GEO, the Local Census Office and Type of Enumeration Area Recode will need to be assigned on the Sample Design File. A Prioritization file will be created for Field Division. This will allow them to prioritize their HUFU and PFU workloads prior to the processing operation.

A. UPDATE SAMPLE DESIGN FILE

After the operational GRF is received from the GEO, updates will need to be made to the Sample Design File. The Local Census Office and the Type of Enumeration Area Recode will be assigned.

On the Sample Design File, update the following variables for all block clusters that remain in sample:

Variable Description	Name
Local Census Office	LCO
Type of Enumeration Area Recode	TEACR
1 = City-Style Address	
2 = Non-City-Style Address	

After all states have been verified, concatenate the 52 separate state Sample Design Files into one file, ACE2000_SDFV1.<mmddyy> .

B. FIELD PRIORITIZATION FILE

After the GEO updates the TEA and Local Census Office information and returns the file to the DSCMO, a file of the block clusters remaining in sample will be created for Field Division. This will be a block cluster level file and contain all block clusters that remain in sample from the 50 states, the District of Columbia and Puerto Rico. The layout of the file is in Appendix G.

Output the following variables for each block cluster that remains in sample on the file:

Variable Description	Name
Regional Office	RO
Local Census Office	LCO
ACE Cluster Number	CLUST
Type of Enumeration Area Recode	TEACR
1 = City-Style Address	
2 = Non-City-Style Address	

cc: DSSD Census 2000 Procedures and Operations Memorandum Series Distribution List
ACE Implementation Team/Statistical Design Team Leaders List
DSSD Sample Design Team
S. Odell (DSSD)
C. Hantman (GEO)
R. Ruiz (GEO)
S. Holt (GEO)
K. Todd (GEO)
S. Hawala (PRED)

Appendix A

Block Cluster Sampling Parameter File Layout

Variable Description	Name	Places
Census Region	REGION	1
Census Division	DIV	2
State code (01-72 = FIPS State Code)	STATE	3-4
Sampling Stratum	SS	5
Target number of block clusters	TCLUST	7-14
Total number of block clusters	NCLUST	16-23
Total number of housing units	NHU	25-32
First stage Take-Every	TE1	34-44
First stage Random Start	RS1	46-56
Indicator for Second-Step of Block Cluster Sampling	I2	58-58
Second-Step Random Start	RS2	60-70
Second-Step Take-Every	TE2	72-82
Clusters in Sample to List	CLUSL	84-91
Housing Units in Sample to List After First Step	INMHUL	93-100
Housing Units in Sample to List After Second Step	NMHUL	102-109

Universe File Layout

Variable Description	Name	Places
State	STATE	1-2
County	COUNTY	3-5
Interim Tract (a.k.a. pseudo-tract)	ITRACT	6-11
Block Number	COLBLOCK	12-16
Blank		17-17
Cluster Number (geography not ACE)	GCLUS	18-22
Blank		23-23
Cluster Size code	CLUSSIZE	24-24
1 = Clusters with 0 HUs		
2 = Clusters with 1 HUs		
3 = Clusters with 2 HUs		
4 = Clusters with between 3 and 5 HUs		
5 = Clusters with between 6 and 9 HUs		
6 = Clusters with between 10 and 19 HUs		
7 = Clusters with between 20 and 29 HUs		
8 = Clusters with between 30 and 79 HUs		
9 = Clusters with 80 or more HUs		
Blank		25-25
Block Area (Sq. Miles)	BAREA	26-33
Blank		34-34
Block Perimeter (Miles)	BPERIM	35-40
Blank		41-41
Block Cluster Area (Sq. Miles)	BCAREA	42-49
Blank		50-50
Block Cluster Perimeter (Miles)	BCPERIM	51-56
Number of HUs in cluster	NHU	57-61
Number of HUs in block	NHUBLOCK	62-66
Block TEA	TEA	67-67
1 = Mailout/Mailback		
2 = Update/Leave		
3 = List/Enumerate		
5 = Rural Update/Enumerate		
6 = Military		
7 = Urban Update/Leave		
8 = Update/Leave to Mailout/Mailback conversions		
9 = Mailout/Mailback to Update/Leave conversions		

TEA Group for Block Cluster	TEABC	68-68
A= Mailout/Mailback or Urban Update/Leave or Update/Leave to Mailout/Mailback conversions		
B= Update/Leave or Rural Update/Enumerate		
C=List/Enumerate		
D=Military		
E=Mailout/Mailback to Update/Leave conversions		
2000 MAF HUs count	NHUM	69-73
' ' Blank if no HU count available		
1990 ACF HUs count	NHU90	74-78
' ' Blank if no HUs count available		
Housing Unit Count Indicator	HUIND	79-79
1 = from 2000 MAF		
2 = from 1990 ACF		
Invisible Boundary Collapse Indicator	INV	80-80
0 = No		
1 = Yes (Collapsing across Invisible Boundary in BC)		
American Indian Country Indicator	AICIND	81-81
0 = No American Indian Country		
1 = American Indian Reservation/trust land		
2 = Tribal jurisdiction statistical area/ Alaska Native Village statistical area/ tribal designated statistical area		
Military Indicator	MILIND	82-82
0 = No Military Area		
1 = Block contains Military Area		
Collapsed Enclosed Block Indicator	CEBI	83-83
0 = Otherwise		
1 = An enclosed block has been forced to collapse		

Blank		84-90

2000 Collection Block Estimated Number of:		
Hawaiian and Pacific Islander Renter	ECOLPIR	91-95
Hawaiian and Pacific Islander Owner	ECOLPIO	96-100
American Indian and Alaska Native Renter	ECOLIR	101-105
American Indian and Alaska Native Owner	ECOLIO	106-110
Asian Renter	ECOLAR	111-115
Asian Owner	ECOLAO	116-120
Hispanic Renter	ECOLHR	121-125
Hispanic Owner	ECOLHO	126-130
Black Renter	ECOLBR	131-135
Black Owner	ECOLBO	136-140
White and Other Renter	ECOLOR	141-145
White and Other Owner	ECOLOO	146-150
Total Renters	ECOLR	151-155
Total Owners	ECOLO	156-160
Total Housing Units	ECOLHU	161-165
Occupied Housing Units	ECOLOHU	166-170
Total People (Non-GQ)	ECOLPOP	171-175
Estimated 1990 urbanicity of the 2000 collection block 1 = Urban Area with 1990 population ≥ 250,000 2 = Other Urban Area 3 = Non-Urban Area	ECOLURB	176-176
Blank		177-180
2000 Collection Block Cluster Estimated Number of:		
Hawaiian and Pacific Islander Renter	ECLUSPIR	181-185
Hawaiian and Pacific Islander Owner	ECLUSPIO	186-190
American Indian and Alaska Native Renter	ECLUSIR	191-195
American Indian and Alaska Native Owner	ECLUSIO	196-200
Asian Renter	ECLUSAR	201-205
Asian Owner	ECLUSAO	206-210
Hispanic Renter	ECLUSHR	211-215
Hispanic Owner	ECLUSHO	216-220
Black Renter	ECLUSBR	221-225
Black Owner	ECLUSBO	226-230
White and Other Renter	ECLUSOR	231-235
White and Other Owner	ECLUSOO	236-240
Total Renters	ECLUSR	241-245
Total Owners	ECLUSO	246-250
Total Housing Units	ECLUSHU	251-255
Occupied Housing Units	ECLUSOHU	256-260
Total People (Non-GQ)	ECLUSPOP	261-265
Blank		266-275

Estimated 1990 urbanicity of 2000 block cluster	ECLUSURB	276-276
1 = Urban Area with 1990 population ≥ 250,000		
2 = Other Urban Area		
3 = Non-Urban Area		
Size Category	SIZECAT	277-277
1 = Small (0-2 HUs)		
2 = Medium (3-79 HUs)		
3 = Large (80+ HUs)		
Number of sampling strata in state	NSSINST	278-278
Sample stratum	SS	279-279
1 = Small		
2 = Medium (non-AIR)		
3 = Large (non-AIR)		
4 = American Indian Reservation		
Blank		280-285
2000 Collection Block Cluster Proportion of Population that is:		
Hawaiian and Pacific Islander Renter	CLUPPIR	286-290
Hawaiian and Pacific Islander Owner	CLUPPIO	291-295
American Indian and Alaska Native Renter	CLUPIR	296-300
American Indian and Alaska Native Owner	CLUPIO	301-305
Asian Renter	CLUPAR	306-310
Asian Owner	CLUPAO	311-315
Hispanic Renter	CLUPHR	316-320
Hispanic Owner	CLUPHO	321-325
Black Renter	CLUPBR	326-330
Black Owner	CLUPBO	331-335
White and Other Renter	CLUPOR	336-340
White and Other Owner	CLUPOO	341-345
Renters	CLUPR	346-350
Owners	CLUPO	351-355
Blank		356-364
Demographic/Tenure group (code)	DTCODE	365-366
Demographic/Tenure group (label)	DTLABEL	367-368
Region	REGION	369-369
Division	DIV	370-370
Blank		371-399

Current Sample Indicator 0 = Not in Sample 1 = In Sample	CSI	400-400
First-Step Block Cluster Sample Indicator	BC1	402-402
First-Step Index Number	INDEX1	404-411
Second-Step Block Cluster Sample Indicator	BC2	413-413
Second-Step Index Number	INDEX2	415-422

Sample Design File Layout

Variable Description	Name	Places
Census Region	REGION	1
Census Division	DIV	2
State code	STATE	3-4
County code	COUNTY	5-7
Local census office	LCO	8-11
Interim Tract (Pseudo Tract)	ITRACT	12-17
Current Sample Indicator	CSI	19
ACE block cluster number	CLUST	21-25
Check Digit	DIGIT	26
Geography block cluster number	GCLUST	28-32
Type of Enumeration Area Recode	TEACR	34
Type of Enumeration Area group	TEAG	36
Number of HUs used for sample design	NHU	37-41
Number of MAF HUs	NHUM	43-47
Number of 1990 HUs	NHU90	49-53
Sampling Stratum	SS	55
1 = Small		
2 = Medium		
3 = Large		
4 = American Indian Reservation		
American Indian Country Indicator	AICIND	56
0 = No American Indian Country		
1 = American Indian Reservation/trust land		
2 = Tribal Jurisdiction Area/ Alaska Native Village Statistical Area/ Tribal Designated Statistical Area		
Demographic/Tenure Group code	DTCODE	57-58
Demographic/Tenure Group label	DTLABEL	59-60
Estimated Urbanicity of block cluster	ECLUSURB	62
1 = Urban Area with population ≥250,000		
2 = Other Urban Area		
3 = Non-Urban Area		
Size Category	SIZCAT	63
1=Small (0-2 hus)		
2=Medium (3-79 hus)		
3=Large (80+ hus)		
Additional space		64-91

First step index number	INDEX1	92-99
Initial block cluster sampling Indicator 1 = Selected	BC1	101
Random Start for initial block cluster sampling	RS1	103-113
Take-every for initial block cluster sampling	TE1	115-125
Second block cluster sampling Indicator 0 = Not Selected, 1 = Selected	BC2	127
Random Start for second block cluster sampling	RS2	129-139
Take-every for second block cluster sampling	TE2	141-151
Unbiased weight after block cluster sampling	WEIGHTBC	153-164

ACE Summary File Layout

There will be one file for each state. This file will have one record per implicit stratum in each sampling stratum.

Variable Description	Name	Location
FIPS State Code	STATE	1:2
Sampling Stratum 1 = Small Block Clusters, 2 = Medium Block Clusters, 3 = Large Block Clusters, 4 = American Indian Reservation (AIR) Block Clusters	SS	4:4
Demographic/Tenure Group Code	DTCODE	6:7
Number of Block Clusters	S_NCLUST	10:17
Number of Housing Units	S_NHU	20:27
Number of Black/Owner People in 1990	S_BOP	30:37
Number of Black/Renter People in 1990	S_BR	40:47
Number of Hispanic/Owner People in 1990	S_HOP	50:57
Number of Hispanic/Renter People in 1990	S_HRP	60:67
Number of Asian/Owner People in 1990	S_AOP	70:77
Number of Asian/Renter People in 1990	S_ARP	80:87
Number of Hawaiian and Pacific Islander/Owner People in 1990	S_POP	90:97
Number of Hawaiian and Pacific Islander/Renter People in 1990	S_PR	100:107
Number of American Indian Reservation/Owner People in 1990	S_AIROP	110:117
Number of American Indian Reservation/Renter People in 1990	S_AIRRP	120:127
Number of American Indian Not on Reservation/Owner People in 1990	S_AICOP	130:137
Number of American Indian Not on Reservation/Renter People in 1990	S_AICRP	140:147
Number of White and Other/Owner People in 1990	S_OOP	150:157
Number of White and Other/Renter People in 1990	S_ORP	160:167

ACE Sample Cluster File Layout

Variable Description	Name	Location
FIPS State Code	STATE	1:2
Local Census Office	LCO	4:7
(To be filled by the GEO)		
FIPS County Code	COUNTY	9:11
ACE Cluster Number	CLUST	13:17
Check Digit	DIGIT	18:18
Sampling Strata	SS	20:20
1 = Small Block Clusters,		
2 = Medium Block Clusters,		
3 = Large Block Clusters,		
4 = American Indian Reservation (AIR) Block Clusters		
Demographic/Tenure Group Code	DTCODE	21:21
2000 Collection Block	BK2K	23:27
Geography Cluster Number from Geography	GCLUST	29:33
Cluster Size Recode from Geography	GSIZE	36:36
Number of Housing Unit in Cluster	NHU	38:42
Number of Housing Units in Block	NHB	44:48
Number of 2000 MAF Housing Units in Block	NHUMB	50:54
Number of 1990 Estimated Housing Units in Block	NHU90B	56:60
Block Type of Enumeration Area Revised	TEABR	62:62
(To be filled by the GEO)		
Type of Enumeration Area Recode	TEACR	64:64
(To be filled by the GEO)		
1 = City-Style Address		
2 = Non-City-Style Address		
Total Persons in the Cluster	NP	66:73

Appendix F

Sample Summary File Layout

Variable Description	Name	Places
Census Region	REGION	1
Census Division	DIV	2
State code (01-72 = FIPS State Code)	STATE	3-4
Number of HUs budgeted for listing in med. and lg. clusters	BLIST	6-13
Target number of clusters in med. and lg. sampling strata	TCLUST	15-18
Target number of clusters in small sampling strata	TCLUST	20-22
Target number of clusters in AIR sampling strata	TCLUST	24-26
Total number of block clusters	NCLUST	28-35
Total number of HUs	NHU	37-44
Expected clusters in sample to list	ECLUST	46-49
Expected HUs in sample to list	EXPHUL	51-58
Additional space		59-80

Clusters in sample to list after 1st step sampling	NCLUSTL1	81-85
Estimated HUs in sample to list after 1st step sampling	NHUL1	87-94
Estimated HUs in sample to list after 1st step sampling in Med & Lg clusters	NHUL1_ML	96-103
Indicator for second step of block cluster sampling 1 = Second step needed, 2 = Second step not needed	I2	105
Clusters in sample to list after 2nd step sampling	NCLUSTL2	107-111
Estimated HUs in sample to list after 2nd step sampling	NHUL2	113-120
Estimated HUs in sample to list after 2nd step sampling in Med & Lg clusters	NHUL2_ML	122-129

Field Prioritization File Layout

Variable Description	Name	Position
Regional Office	RO	1:2
Local Census Office	LCO	4:7
ACE Cluster Number	CLUST	9:13
Number of Hus for Sample Design	DIGIT	14:14
Number of HUs for Sample Design.	NHU	16:23
TEA Revised Code 1 = City-Style Address 2 = Non-City-Style Address	TEACR	25:25

Appendix H

Sample Size Input File Layout

Variable Description	Name	Places
Census Region	REGION	1
Census Division	DIV	2
State	STATE	3-4
Number of housing units budgeted for listing	BLIST	6-13
Target Clusters for small clusters	TCLUSTS	15-17
Target Clusters for medium and large clusters	TCLUST	19-22
Target Clusters for AIR	TCLUST	24-26
Number of sampling stratum in state	NSSINST	28
First ACE block cluster number	CSTART	30-34

State Budgeted Number of Housing Units for Listing and First-Step Target Block Cluster Sample Sizes

State	Region	Division	Number of Strata	Budgeted Listing	AIR Cluster Target	Small Cluster Target	Medium and Large Cluster Target
Alabama*	3	6	3	25,347	0	116	417
Alaska	4	9	4	27,196	1	20	334
Arizona	4	8	4	48,451	110	86	492
Arkansas	3	7	3	24,744	0	90	494
California	4	9	4	284,076	14	184	2,753
Colorado	4	8	4	37,965	2	83	479
Connecticut*	1	1	3	30,039	0	20	377
Delaware	3	5	3	21,610	0	20	413
DC	3	5	3	53,369	0	20	384
Florida	3	5	4	62,845	1	145	520
Georgia*	3	5	3	37,384	0	154	399
Hawaii	4	9	3	45,059	0	20	300
Idaho	4	8	4	19,157	6	54	412
Illinois	2	3	3	31,571	0	185	430
Indiana	2	3	3	15,925	0	140	275
Iowa*	2	4	3	14,108	0	147	300
Kansas	2	4	4	16,281	1	193	300
Kentucky	3	6	3	29,621	0	96	447
Louisiana*	3	7	3	37,378	0	65	595
Maine	1	1	4	16,572	1	38	309
Maryland	3	5	3	41,107	0	36	368
Massachusetts*	1	1	3	27,255	0	38	375
Michigan	2	3	4	24,128	5	122	379
Minnesota	2	4	4	19,091	10	141	300
Mississippi	3	6	4	19,990	3	81	402
Missouri	2	4	3	19,807	0	162	300
Montana	4	8	4	17,969	24	67	420
Nebraska	2	4	4	14,177	3	142	300
Nevada	4	8	4	63,031	5	46	468
New Hampshire	1	1	3	21,128	0	25	307
New Jersey**	1	2	3	37,394	0	39	461
New Mexico	4	8	4	32,242	70	108	481
New York	1	2	4	143,949	5	143	1,261
North Carolina	3	5	4	26,717	4	143	400
North Dakota	2	4	4	15,738	12	121	300
Ohio	2	3	3	30,790	0	132	421

State	Region	Division	Number of Strata	Budgeted Listing	AIR Cluster Target	Small Cluster Target	Medium and Large Cluster Target
Oklahoma	3	7	4	25,328	8	142	426
Oregon	4	9	4	20,577	3	86	320
Pennsylvania	1	2	3	34,920	0	180	585
Rhode Island*	1	1	3	23,557	0	20	373
South Carolina*	3	5	3	26,709	0	95	422
South Dakota	2	4	4	14,227	27	106	300
Tennessee	3	6	3	30,255	0	133	433
Texas	3	7	4	176,234	1	349	1,945
Utah	4	8	4	32,777	7	38	478
Vermont	1	1	3	17,009	0	21	300
Virginia*	3	5	3	37,114	0	98	371
Washington	4	9	4	26,832	17	73	332
West Virginia	3	5	3	17,557	0	46	300
Wisconsin	2	3	4	14,470	10	119	275
Wyoming	4	8	4	18,293	5	72	418
United States				1,949,070	355	5000	24,601
Puerto Rico	0***	0***	3	46,700	0	96	480

* States contain AIR population, but not AIR sampling stratum. AIR people will be given a chance of selection in the general state sample.

** New Jersey AIR reservations had no population in 1990.

*** Puerto Rico is an outlying area. Because of this it has no region or division code. Therefore, we assign the code 0.

ACE Block Cluster Number Allocation

Division	State	ACE Block Cluster Number
New England	Connecticut	11001-11999
	Maine	12001-12999
	Massachusetts	13001-13999
	New Hampshire	14001-14999
	Rhode Island	15001-15999
	Vermont	16001-16999
	Puerto Rico	17001-17999
Mid-Atlantic	New Jersey	21001-21999
	New York	22001-24999
	Pennsylvania	25001-25999
South Atlantic	Delaware	31001-31999
	DC	32001-32999
	Florida	33001-33999
	Georgia	34001-34999
	Maryland	35001-35999
	North Carolina	36001-36999
	South Carolina	37001-37999
	Virginia	38001-38999
	West Virginia	39001-39999
East South Central	Alabama	41001-41999
	Kentucky	42001-42999
	Mississippi	43001-43999
	Tennessee	44001-44999
West South Central	Arkansas	51001-51999
	Louisiana	52001-52999
	Oklahoma	53001-53999
	Texas	54001-57999
East North Central	Illinois	61001-61999
	Indiana	62001-62999
	Michigan	63001-63999
	Ohio	64001-64999
	Wisconsin	65001-65999

Division	State	ACE Block Cluster Number
West North Central	Iowa	71001-71999
	Kansas	72001-72999
	Minnesota	73001-73999
	Missouri	74001-74999
	Nebraska	75001-75999
	North Dakota	76001-76999
	South Dakota	77001-77999
Mountain	Arizona	81001-81999
	Colorado	82001-82999
	Idaho	83001-83999
	Montana	84001-84999
	Nevada	85001-85999
	New Mexico	86001-86999
	Utah	87001-87999
	Wyoming	88001-88999
Pacific	Alaska	91001-91999
	California	92001-96999
	Hawaii	97001-97999
	Oregon	98001-98999
	Washington	99001-99999

Appendix K

Double-Add-Double Check-digit Algorithm

1. Set working count (WC) to 0.

2. Set working data (WD) to the binary value of the input string

3. Look at the rightmost digit of WD
   (i.e. start with the units digit which is MOD(WD,10)) if its 0 add 0 to WC
   if its 1 add 2 to WC
   if its 2 add 4 to WC
   if its 3 add 6 to WC
   if its 4 add 8 to WC
   if its 5 add 1 to WC
   if its 6 add 3 to WC
   if its 7 add 5 to WC
   if its 8 add 7 to WC
   if its 9 add 9 to WC
   (the added value is 'double' the input - if the result is 10+, the tens digit and units digit are added)

4. Add the second rightmost digit of WD to WC
   (i.e. start with the tens digit which is MOD(WD/10,10))

5. Shift WD 2 digits to the right (i.e. WD = WD/100)

6. While WD is not zero, repeat from step 3.

7. set WC to MOD10 of WC

8. If WC is 1 to 9 then set the CHECKDIGIT to 10-WC.
   If WC is 0 then set the CHECKDIGIT to 0.

Example:

the CHECKDIGIT for 123456 is 6, because
WC = 'double'6 + 5 + 'double'4 + 3 + 'double'2 + 1 = 3 + 5 + 8 + 3 + 4 + 1 = 24 so the
CHECKDIGIT is 10-MOD(24,10) = 6.

UNITED STATES DEPARTMENT OF COMMERCE
Bureau of the Census
Washington, DC 20233-0001

MASTER FILE

March 30,1999

DSSD CENSUS 2000 PROCEDURES AND OPERATIONS MEMORANDUM SERIES R-4

MEMORANDUM FOR Dennis Stoudt
Assistant Division Chief, Processing Systems
Decennial Systems and Contracts Management Office

From: Donna Kostanich DK
Assistant Division Chief, Sampling and Estimation
Decennial Statistical Studies Division

Prepared by: Randy ZuWallack RZ
Sample Design Team
Decennial Statistical Studies Division

Subject: Accuracy and Coverage Evaluation (ACE) Survey: Sample Summary File and Sample Design File Documentation

This memorandum documents the layout of two files that will be continuously updated during the sample selection for the Census 2000 Accuracy and Coverage Evaluation (ACE). The first is the ACE Sample Summary File, which will contain cluster and housing unit totals at the state level. This file will aid in the monitoring of the sampling procedures by providing expected and actual results which will then be compared to identify extreme differences. Attachment A contains a file layout for the Sample Summary File. The second file is the ACE Sample Design File. This file tracks the path that each block cluster travels during the ACE sampling procedures. The Sample Design File contains categorical variables corresponding to each procedure as well as parameters and housing unit totals. In addition, sampling weights will be assigned based on the final path each cluster follows during the ACE sampling operations. Attachment B contains a file layout for the Sample Design File. Together the Sample Summary File and Sample Design File will document the history of the ACE design and serve as a reference during evaluations and estimation.

The creation of the Sample Summary File will occur following the creation of the Universe File[1]. The Sample Design File will be created following the block cluster sampling[2]. The Sample Summary File and Sample Design File will be updated in the specifications for each of the ACE sampling procedures, which include the initial ACE block cluster sampling, the ACE block cluster reduction, small block subsampling, large block subsampling, and E-sample identification. Although the Sample Summary File and Sample Design File will be updated following each of these processes, the layout for these files will be documented in this specification. A source code is assigned to each variable indicating where in the processing the variable is first encountered. These source codes are listed following each file layout. For information not foreseen as being required for the sampling procedures, space will be left for additions to the files. This space will be filled as necessary following each process, and will be documented in the specification for that process. At the conclusion of all ACE sampling operations, the final layout for the Sample Summary File and Sample Design File will be documented.

For questions concerning the Sample Design File or the Sample Summary File, contact Deborah Fenstermaker 301-457-4195 or Randy ZuWallack 301-457-1963.

cc: DSSD Census 2000 Procedures and Operations Memorandum Series Distribution List
ACE Implementation Team/Statistical Design Team Leaders List
Sample Design Team

---

[1]Memorandum from Kostanich to Stoudt, "Accuracy and Coverage Evaluation Survey: Universe File and Sampling Parameter File Specification", March 1999.

[2]Memorandum from Kostanich to Stoudt, "Accuracy and Coverage Evaluation Survey: Block Cluster Sample Selection", March 1999.

Sample Summary File

The Sample Summary File contains one record for each of the 50 states, the District of Columbia and Puerto Rico for a total of 52 records. The initial version of the file, which will be created following the creation of the Universe File, is called ACE2000_SSFV1.<mmddyy>. The extension <mmddyy> is the date the file is created (i.e. 123199 is the extension for a file created on December 31, 1999). For each subsequent update to the file, the version number will increase by one (i.e. ACE2000_SSFV2.<mmddyy>, ACE2000_SSFV3.<mmddyy>). The layout for the Sample Summary File is as follows:

Variable Description	Name	Places	Source
Census Region	REGION	1	UN
Census Division	DIV	2	UN
State code (01-72 = FIPS State Code)	STATE	3-4	UN
Number of HUs budgeted for listing in med. and lg. clusters	BLIST	6-13	UN
Target number of clusters in small sampling strata	TCLUSTS	15-17	UN
Target number of clusters in med. and lg. sampling strata	TCLUST	19-22	UN
Target number of clusters in AIR sampling strata	TCLUSTA	24-26	UN
Total number of block clusters	NCLUST	28-35	BC
Total number of HUs	NHU	37-44	BC
Expected clusters in sample to list	ECLUSTL	46-49	UN
Expected HUs in sample to list	EXPHUL	51-58	UN
Additional space		59-80	
Clusters in sample to list after 1st step sampling	NCLUSTL1	81-85	CS
Estimated HUs in sample to list after 1st step sampling	NHUL1	87-94	CS
Estimated HUs in sample to list after 1st step sampling in Med & Lg clusters	NHUL1_ML	96-103	
Indicator for second step of block cluster sampling 1 = Second step needed, 2 = Second step not needed	I2	105	CS
Clusters in sample to list after 2nd step sampling	NCLUSTL2	107-111	CS
Estimated HUs in sample to list after 2nd step sampling	NHUL2	113-120	CS
Estimated HUs in sample to list after 2nd step sampling in Med & Lg clusters	NHUL2_ML	122-129	CS
Additional space		130-150	
Preliminary Number of HUs on Independent List	NHUILLP	151-158	AR
Number of Housing Units On the DMAF	NHUDMAF	160-167	AR
Additional space reserved for ACE reduction		168-270	
Number of HUs on Independent List	NHUILL	271-278	SB
Expected number of clusters selected for ACE	ECLUST	280-284	SB
Expected number of Independent List HUs for ACE	EHUIL	286-293	SB
Number of clusters selected for ACE	NCLUST	295-299	SB
Number of Independent List HUs for ACE	NHUIL	301-308	SB
Additional space		309-330	

Attachment A

Variable Description	Name	Places	Source
Number of HUs on the Preliminary Enhanced List	NHUEL	331-338	LB
Number of ACE HUs on the Preliminary Enhanced List	NHUELA	340-347	LB
Number of non-ACE HUs on the Preliminary Enhanced List	NHUELN	349-346	LB
Expected number of HUs for interview	EHUINT	358-365	LB
Expected number of ACE HUs for interview	EHUINTA	367-374	LB
Expected number of non-ACE HUs for interview	EHUINTN	376-383	LB
Number of HUs for interview	NHUINT	385-392	LB
Number of ACE HUs for interview	NHUINTA	394-401	LB
Number of non-ACE HUs for interview	NHUINTN	403-410	LB
Additional space		411-430	
Number of CUF HUs	NHUCUF	431-438	ES
Number of CUF HUs in block cluster with an ESPS code of 1	NHUCUF1	440-447	ES
Number of CUF HUs in block cluster with an ESPS code of 2	NHUCUF2	449-456	ES
Expected number of E-sample HUs	EHUES	458-465	ES
Expected number of E-sample HUs with an ESPS code of 1	EHUES1	467-474	ES
Expected number of E-sample HUs with an ESPS code of 2	EHUES2	470-483	ES
Number of E-sample HUs	NHUES	485-492	ES
Number of E-sample HUs with an ESPS code of 1	NHUES1	494-501	ES
Number of E-sample HUs with an ESPS code of 2	NHUES2	503-510	ES
Additional Space		511-600	

Source Codes

AR: ACE Reduction
BC: Block Clustering
CS: Block Cluster Sampling
ES: E-sample Identification
LB: Large Block Subsampling
SB: Small Block Subsampling
UN: Universe File Creation

Attachment B

Sample Design File

The Sample Design File contains one record per block cluster selected during the initial block cluster sampling. If the block clusters falls out of sample during the second step of sampling or during small block subsampling, the remaining variables will be left blank. The initial version of the file, which will be created following the initial block cluster selection, is called ACE2000_SDFV1.<mmddyy>. For each subsequent update to the file, the version number will increase by one (i.e. ACE2000_SDFV2.<mmddyy>, ACE2000_SDFV3.<mmddyy>). The layout for the Sample Design File is as follows:

Variable Description	Name	Places	Source
Census Region	REGION	1	UN
Census Division	DIV	2	UN
State code	STATE	3-4	UN
County code	COUNTY	5-7	UN
Local census office	LCO	8-11	CS
Interim Tract (Pseudo Tract)	ITRACT	12-17	BC
Current Sample Indicator	CSI	19	UO
ACE block cluster number	CLUST	21-25	CS
Check Digit	DIGIT	26	CS
Geography block cluster number	GCLUST	28-32	BC
Type of Enumeration Area Recode	TEACR	34	CS
Type of Enumeration Area group	TEAG	36	BC
Number of HUs used for sample design	NHU	37-41	BC
Number of MAF HUs	NHUM	43-47	BC
Number of 1990 HUs	NHU90	49-53	BC
Sampling Stratum 1 = Small 2 = Medium 3 = Large 4 = American Indian Reservation	SS	55	UN
American Indian Country Indicator 0 = No American Indian Country 1 = American Indian Reservation/trust land 2 = Tribal Jurisdiction Area/ Alaska Native Village Statistical Area/ Tribal Designated Statistical Area	AICIND	56	BC
Demographic/Tenure Group code	DTCODE	57-58	UN
Demographic/Tenure Group label	DTLABEL	59-60	UN
Estimated Urbanicity of block cluster 1 = Urban Area with population ≥250,000 2 = Other Urban Area 3 = Non-Urban Area	ECLUSURB	62	UN

Attachment B

Variable Description	Name	Places	Source
Size Category	SIZCAT	63	UN
1=Small (0-2 hus)			
2=Medium (3-79 hus)			
3=Large (80+ hus)			
Additional space		64-91	
First step index number	INDEX1	92-99	CS
Initial block cluster sampling Indicator	BC1	101	CS
1 = Selected			
Random Start for initial block cluster sampling	RS1	103-113	UN
Take-every for initial block cluster sampling	TE1	115-125	UN
Second block cluster sampling Indicator	BC2	127	CS
0 = Not Selected, 1 = Selected			
Random Start for second block cluster sampling	RS2	129-139	CS
Take-every for second block cluster sampling	TE2	141-151	CS
Unbiased weight after block cluster sampling	WEIGHTBC	153-164	CS
Additional space		165-175	
Preliminary Number of HUs on the Independent List	NHUILP	176-180	AR
Number of Housing Units On the DMAF	NHUDMAF	182-186	AR
Additional space reserved for ACE reduction		187-277	
Unbiased weight after ACE reduction	WEIGHTAR	278-289	AR
Additional space		290-300	
Number of HUs on the Independent List	NHUIL	301-305	SB
Independent List Cluster Category	ILCC	307	SB
Small Block Subsampling Indicator	SB	308	SB
0 = Not Selected, 1 = Selected			
Random Start for Small Block subsampling	RSSB	310-320	SB
Take-every for Small Block subsampling	TESB	322-332	SB
Unbiased weight for ACE cluster	WEIGHTC	334-345	SB
Additional space		346-370	
Relisted Block Cluster Flag	RELIST	371	LB
0 = Not Relisted, 1 = Relisted			
Number of total hus on the EL in block cluster	NHUEL	373-377	LB
Number of ACE hus on the EL in cluster	NHUELA	379-383	LB
Number of non-ACE hus on the EL in cluster	NHUELN	385-389	LB
Enhanced List Cluster Category	ELCC	391	LB
1 = NHUELI< 80 hus, 2 = NHUELI ≥ 80 hus			
Random Start for Large Block subsampling	RSLB	393-403	LB
Take-every for Large Block subsampling	TELB	405-415	LB
Number of Segments per block cluster	NSEG	417-418	LB
Number of selected segments	NSEGSAM	420-421	LB
Day of Arrival	DAY	423-424	LB
Daily Cluster Order Number	DCON	426-429	LB

Attachment B

Variable Description	Name	Places	Source
Final Cluster Order Number	CON	431-434	LB
Non-ACE Subsampling Flag	NISUB	436	LB
Number of total hus for interview in block cluster	NINT	438-442	LB
Number of ACE hus for interview in block cluster	NINTA	444-448	LB
Number of non-ACE HUs for interview	NINTN	450-454	LB
Unbiased weight for P-sample HUs	WEIGHTP	456-467	LB
Additional space		468-490	
Number of CUF HUs in block cluster with an ESPS code of 1	NHUCUF1	491-495	ES
Number of CUF HUs in block cluster with an ESPS code of 2	NHUCUF2	497-501	ES
Number of CUF HUs in block cluster	NHUCUF	503-507	ES
Number of CUF HUs in selected segments with an ESPS code of 1	NHUCUFS1	509-513	ES
Number of CUF HUs in selected segments with an ESPS code of 2	NHUCUFS2	515-519	ES
Number of CUF HUs in selected segments of a block cluster	NHUCUFS	521-525	ES
E-Sample Identification cluster category	EICC	527	ES
1 = NHUCUF < 80			
2 = NHUCUF ≥ 80 and NHUCUFS < 80			
3 = NHUCUF ≥ 80 and NHUCUFS ≥ 80			
4 = NHUCUF ≥ 80 and RELIST = 1			
5 = NHUCUF ≥ 80 and List/Enumerate			
Random Start for E-sample subsampling	RSES	529-539	ES
Take-every for E-sample subsampling	TEES	541-551	ES
Number of E-sample HUs in block cluster with an ESPS code of 1	NHUES1	553-557	ES
Number of E-sample HUs in block cluster with an ESPS code of 2	NHUES2	559-563	ES
Number of E-sample HUs in block cluster	NHUES	565-569	ES
Unbiased weight for E-sample HUs with an ESPS code of 1	WEIGHTE1	571-582	ES
Unbiased weight for E-sample HUs with an ESPS code of 2	WEIGHTE2	584-595	ES
Additional Space		596-620	
Trimmed weight for P-sample HUs	TRIMWTP	621-632	WT
Trimmed weight for E-sample HUs with an ESPS code of 1	TRIMWTE1	634-645	WT
Trimmed weight for E-sample HUs with an ESPS code of 2	TRIMWTE2	647-658	WT
Additional Space		659-750	

Source Codes

AR: ACE Reduction
BC: Block Clustering
CS: Block Cluster Sampling
ES: E-sample Identification
LB: Large Block Subsampling
SB: Small Block Subsampling
UN: Universe File Creation
UO: Updated for each operation
WT: Weight Assignment

UNITED STATES DEPARTMENT OF COMMERCE
Bureau of the Census
Washington, DC 20233-0001

MASTER FILE

March 30,199

DSSD CENSUS 2000 PROCEDURES AND OPERATIONS MEMORANDUM SERIES R- 5

MEMORANDUM FOR Dennis Stoudt
Assistant Division Chief, Processing Systems
Decennial Systems and Contracts Management Office

From: Donna Kostanich DK
Assistant Division Chief, Sampling and Estimation
Decennial Statistical Studies Division

Prepared by: Matt Salganik MS and Randy ZuWallack RZ
Decennial Statistical Studies Division

Subject: Accuracy and Coverage Evaluation (ACE) Survey: Universe File and Block Cluster Sampling Parameter File Specification

I. Introduction

This specification describes how to create the Universe File and the Block Cluster Sampling Parameter File for the selection of the initial ACE sample clusters. The Universe File contains sampling information for all the 2000 block clusters. The initial ACE sample block clusters will be selected from the Universe File using the Block Cluster Sampling Parameter File. One Universe File will be created for each of the 50 states, the District of Columbia, and Puerto Rico. The resulting Universe and Block Cluster Sampling Parameter Files will be input to the initial block cluster sampling operation which is specified separately.

The initial sample of block clusters is allocated to states according to the previously planned Integrated Coverage Measurement (ICM) 750,000 housing unit design. This sample will be provided to the Field Division for independent listing. The results from the independent listing will be used to select a reduced sample for ACE of approximately 300,000 housing units. Plans are underway to redesign the ICM into the ACE. Requirements and details of the ACE design are not known at this time.

Before the ACE universe can be created for each state, the block clustering operation must be completed and approved. The Geography Division (GEO) is producing block clusters for each state on a flow-basis. As the Decennial Statistical Studies Division (DSSD) reviews the block clustering for each state, the DSSD will notify the GEO that a state has been approved, and the GEO will deliver the Block Cluster File for the state to the Decennial Systems and Contracts Management Office (DSCMO). This state flowing process will continue through the creation of the ACE Universe File and Block Cluster Sampling Parameter File. After the DSSD reviews and approves the Universe File and Block Cluster Sampling Parameter File for a state, the block cluster sampling will occur for that state.

These specifications should be used to flowchart the process, to generate further discussion on requirements, to identify and finalize the record layouts of input and output files, and to write computer software to implement the methodology. During and after a testing phase, it is likely that changes to the specifications will be necessary.

The sections of this specification are ordered as follows:

- The second section states the assumptions and definitions of the file creation process.
- The third section describes the creation of the Universe File.
- The fourth section describes the creation of the Block Cluster Sampling Parameter File.
- The fifth section describes the inputs into the entire process.
- The sixth section describes the outputs of the entire process.

Any comments or questions should be directed to Randy ZuWallack (x1963), Matt Salganik (x3636), or Debbie Fenstermaker (x4195).

II. Assumptions/Definitions

A. All 50 states plus the District of Columbia will use the same set of criteria for assigning block clusters to the demographic and tenure groups.

B. Block clusters are classified into three size categories based on the number of housing units: small (0 to 2), medium (3 to 79), and large (80+).

C. A separate sample of small block clusters will be drawn from each state.

D. Large block clusters will be selected at a higher rate than mediums to complement a subsequent housing unit subsampling operation in large blocks.

E. A separate American Indian Reservation (AIR) sample will be selected.

F. Remote Alaska will not be part of the ACE universe and therefore will not be sampled.

III. Creating the Universe File

The Universe File will be a 2000 collection block level file with cluster level information for sorting and sampling. It will also be the source of information for calculating the sample parameters in the next section of this specification. A separate Universe File will be created for each state. These Universe Files will be the input into the block cluster sampling operation specified in a different memorandum.

Building the Universe File requires two major operations. First, the data from the 1990 census is merged with the 2000 collection blocks. Then each block cluster is categorized into several groups which will facilitate assigning clusters to sampling strata and sorting before sample selection. Process each state separately as specified below.

A. Assigning 1990 Data to 2000 Collection Blocks

The purpose of this section of the specification is to explain the process for using information about the 1990 tabulation blocks to estimate the characteristics of 2000 collection blocks.

The relationships between the 1990 tabulation blocks and the 2000 collection blocks are complex. In some cases one 1990 tabulation block is associated with many 2000 collection blocks. In some cases it is the other way around -- several 1990 tabulation blocks are associated with one 2000 collection block. It is also possible for one 1990 tabulation block to be associated with one and only one 2000 collection block. The information about these associations is contained in the Block Equivalency Files that the GEO has created.

The method we will use accounts for these changes in the block definitions and proportionally assigns the 1990 population and housing unit (HU) counts to the 2000 collection blocks. It will also assign the 1990 Urban/Rural and Urban Area Size Codes to the 2000 collection blocks. This method is detailed below and is accompanied by an example.

The next steps will create the Merged Data File which is a collection of the data needed to create the Universe Files.

1. Gathering 1990 Census Data

The basis for this operation is the Block Equivalency Files (see Attachment A for a layout). The GEO created one file for each county. Each record on the Block Equivalency File represents a 1990 tabulation block/2000 collection block association. The following steps place data from the Block Cluster File and the 1990 Hundred Percent Edited Detail File (HEDF) onto the Block Equivalency File.

a. For each state, append the county Block Equivalency Files together to create a state file. From this point on, this file will be referred to as the Merged Data File.

b. Sort the Merged Data File by county and 2000 collection block number.

c. Sort the Block Cluster File (for a layout see Attachment B) by county and 2000 collection block number.

d. For each 2000 collection block in the Merged Data File, obtain the HU count and geography cluster number for that block from the Block Cluster File. There are several HU counts on the Block Cluster File. Use the count found in the 'Number of HUs in Block' field. Append this count and the cluster number to every record involving the 2000 collection block in the Merged Data File. It is important to note that each collection block may have several records because of the nature of the Merged Data File. Each appearance of a collection block should have an HU count and cluster number appended to it.

   Because Remote Alaska will not be sampled, it will not be included on the Block Cluster File. Similarly, water blocks will not be on the Block Cluster File. However, both of these block types will be included on the Merged Data File. Merging the two files will result in records containing blank values. Delete these records from the Merged Data File.

e. Sort the blocks within the Merged Data File by county, 1990 tract, and 1990 tabulation block number and suffix.

f. For each 1990 tabulation block plus suffix, use race, Hispanic origin, and tenure variables on the 1990 HEDF to tally the number of people living in regular HUs in the following 12 demographic/tenure groups. Do not include people living in Group Quarters (GQ). See Attachment C for definitions of each group. Demographic information does not exist for Puerto Rico, so set these 12 values to zero. Append these values to every record corresponding to the 1990 tabulation block plus suffix in the Merged Data File.

Hawaiian and Pacific Islander Renter
Hawaiian and Pacific Islander Owner
American Indian and Alaska Native Renter
American Indian and Alaska Native Owner
Asian Renter
Asian Owner
Hispanic Renter
Hispanic Owner
Black Renter
Black Owner
White and Other Renter
White and Other Owner

Hispanics can be of any race, however we want these groups to be mutually exclusive. So, place all Hispanics in the Hispanic group regardless of race.[1] For example, if a person is a Hispanic Asian Renter, then she is classified as a Hispanic Renter in this process

g. Calculate the total number of people living in owned HUs and total number of people living in rented HUs in each 1990 tabulation block plus suffix. These two values can be calculated for Puerto Rico using the H4 variable on the HEDF (see Attachment C for more information). Append these values to every record corresponding to the 1990 tabulation block plus suffix in the Merged Data File.

Total Renters
Total Owners

h. Calculate the total number of regular HUs and the total number of occupied regular HUs in each 1990 tabulation block plus suffix (see Attachment C for definitions). This information may be used later by the Long Form Sample Design and Estimation Team. Do

---

[1]Note that Hispanic American Indians and Hispanic Alaska Natives living on American Indian reservations, trust lands, tribal jurisdiction statistical areas, tribal designated statistical areas, and Alaska native village statistical areas may be classified as American Indian and Alaska Native for estimation purposes. For sampling purposes, we will treat Hispanic American Indians and Hispanic Alaska Natives as Hispanic.

not include any people living in GQs in any of these calculations. Append these values to every record corresponding to the 1990 tabulation block plus suffix in the Merged Data File.

Total Housing Units
Occupied Housing Units

i. Calculate the total number of people living in HUs in each 1990 tabulation block plus suffix. Append this value to every record corresponding to the 1990 tabulation block plus suffix in the Merged Data File.

Total People (Non-GQ)

j. From the 1990 HEDF, get the geography codes listed below for each 1990 tabulation block plus suffix. For 1990 tabulation blocks containing zero housing units and zero people, obtain the geography codes from the 1990 Geographic Reference File.

1990 Urban/Rural
1990 Urban Area Size Code

Append these codes to every record corresponding to the 1990 tabulation block plus suffix in the Merged Data File. See Attachment D for a list of the possible values of the Urban Area Size Code.

2. Assign 1990 Tabulation Block Information to 2000 Collection Blocks

The Merged Data File now has information about the 1990 tabulation blocks. This section describes the procedure used to create estimates for the 2000 collection blocks. An example displayed in Tables 1 and 2 illustrates this process.

a. At this point the Merged Data File should still be sorted by county, 1990 census tract, 1990 tabulation block and suffix. For each 1990 tabulation block plus suffix, sum the number of HUs in all 2000 collection blocks associated with that tabulation block. Use the 'Number of HUs in Block' field that was taken from the Block Cluster File. Append this total to each record where the tabulation block appears in the Merged Data File.

A 2000 collection block would be considered associated with a 1990 tabulation block if a portion of the collection block falls within the boundaries of the tabulation block.

b. For each record in the Merged Data File do the following:

i. Create a proportion, P, defined to be the number of HUs in the 2000 collection block divided by the total HUs associated with the 1990 tabulation block. Round P to 6 decimal places (0.0000005 rounds up to 0.000001). If the Total 2000 HUs associated with a 1990 tabulation block is equal to zero then P is undefined. In these cases set P = 0.

$$P = \frac{2000 \textit{ collection block HUs}}{\textit{Total } 2000 \textit{ HUs associated with } 1990 \textit{ tabulation block}}$$

ii. Multiply this proportion (P) by the number of people in each of the 12 demographic/tenure groups, the total number of owners and renters, the number of HUs, and the number of occupied HUs in the 1990 tabulation block.

iii. Append these 16 counts (the products computed in ii.) to the Merged Data File. Round these figures to four decimal places (0.00005 rounds up to 0.0001).

c. Sort the Merged Data File by county and 2000 collection block.

d. For each 2000 collection block, sum each demographic/tenure group, the estimated numbers of owners and renters, and the estimated number of HUs and occupied HUs across all records for that 2000 collection block. Round to the nearest integer (0.5 rounds to 1). These summations are the estimated population counts for the 2000 collection blocks. Append these 16 estimates to each record for the 2000 collection block in the Merged Data File.

Now the Merged Data File contains two HU counts for each 2000 collection block. One is based on the HU count from block clustering and will be referred to as the HU count. The other is based on the 1990 tabulation block information and will be referred to as the estimated HU count. This second count will be used later by the Long Form Sample Design and Estimation Team.

e. Calculate the total number of estimated people in each 2000 collection block by summing the estimated number of owners and the estimated number of renters. Append this value to each record for the 2000 collection block in the Merged Data File.

The following example depicts a site containing five 1990 tabulation blocks and five 2000 collection blocks. The example is limited to two demographic/tenure groups for ease of demonstration.

Table 1 shows how to assign population characteristics to 2000 collection block parts (steps 2a and 2b). Table 2 shows how to sum across these 2000 collection block parts to get 2000 collection block totals (step 2d).

Table 1. Assigning Population Characteristics to 2000 Collection Block Parts

1990 Tabulation Block	2000 Collection Block	Housing units in Collection block	1990 Tabulation Block Persons		Total housing units in all Collection Blocks associated with Tabulation Block	P	Persons assigned to 2000 collection block	
			Black Owner	Asian Renter			Black Owner	Asian Renter
117	14157	36	60	100	48	36/48	45	75
117	14158	12				12/48	15	25
118	14157	36	144	0	72	36/72	72	0
118	14158	12				12/72	24	0
118	14167	24				24/72	48	0
119	14157	36	100	50	36	36/36	100	50
120	14159	48	5	30	48	48/48	5	30
121	14158	12	120	12	84	12/84	17.1429	1.7143
121	14163	72				72/84	102.8571	10.2857

Table 2. Summing of 2000 Collection Block Parts

2000 Collection Block		1990 Tabulation Block 117	118	119	120	121	Total
14157	Black Owners	45	72	100	0	0	217
	Asian Renters	75	0	50	0	0	125
14158	Black Owners	15	24	0	0	17.1429	56
	Asian Renters	25	0	0	0	1.7143	27
14159	Black Owners	0	0	0	5	0	5
	Asian Renters	0	0	0	30	0	30
14163	Black Owners	0	0	0	0	102.8571	103
	Asian Renters	0	0	0	0	10.2857	10
14167	Black Owners	0	48	0	0	0	48
	Asian Renters	0	0	0	0	0	0

3. Assign 1990 Geography to 2000 Collection Blocks

The Merged Data File has the 1990 Urban/Rural and 1990 Urban Area Size variables for each 1990 tabulation block. This section describes how to assign these 1990 geographical variables to the 2000 collection blocks.

a. Sort the blocks within the Merged Data File by county, 1990 tract, and 1990 tabulation block number and suffix.

b. For each 1990 tabulation block record on the Merged Data File create a new variable, 1990 tabulation block urbanicity, using the variable 1990 Urban Area Size Code (see Attachment D for possible values of this variable) and 1990 Urban/Rural. This new variable will have three values:

1 = Urban Area with 1990 population ≥ 250,000
2 = Other Urban Area
3 = Non-Urban Area

Use the following algorithm to create 1990 tabulation block urbanicity:

If 1990 Urban Area Size ≥ 19 then
  1990 tabulation block urbanicity = 1
else if 1990 urban/rural = 0 then
  1990 tabulation block urbanicity = 2
else 1990 tabulation block urbanicity = 3

c. Sort the Merged Data File by county and 2000 collection block.

d. For each 2000 collection block, compare the number of people in all the 1990 tabulation blocks associated with the 2000 collection block. Choose the 1990 tabulation block with the most people. Assign that tabulation block's urbanicity to the estimated urbanicity of the 2000 collection block. Make this change for all the 1990 tabulation block/2000 collection block associations that involve the 2000 collection block.

If there is a tie for having the most people, then compare the 1990 tabulation block urbanicity values of all blocks involved in the tie. Assign the lowest value (i.e. closest to 0) to the estimated 1990 urbanicity of the 2000 collection block. Again, make this change for all 1990 tabulation block/2000 collection block associations that involve the 2000 collection block.

4. Produce Merged Data File Verification Outputs

Provide the DSSD access to the Merged Data Files for verification. The DSSD will verify the creation of the Merged Data File for eight selected counties. See Attachment E for a list of counties. See Attachment F for the layout of the Merged Data File.

B. Create the Universe File

At this point we have all the data we need on the Merged Data File and we will begin using that information to create the Universe File -- a block level file with block and cluster information. The basis of this file will be the Block Cluster File (see Attachment B for a layout). The following are the steps to create the Universe File which include assigning stratification and sampling variables.

1. Calculate Cluster Level Variables

The Merged Data File has 2000 collection block variables. These variables will be transferred to the Block Cluster File and then summarized to the cluster level.

a. Sort the Merged Data File by county and 2000 collection block.

b. Sort the Block Cluster File by county and 2000 collection block. From this point on, the Block Cluster File will be referred to as the Universe File.

c. For each 2000 collection block, append the estimated number of people in each of the 12 demographic/tenure groups, the two tenure groups, and the estimated number of HUs and occupied HUs to the 2000 collection block record in the Universe File. These counts can be found in the Merged Data File.

   The Universe File now contains two HU counts for each 2000 collection block. One was used to do the block clustering and will be referred to as the HU count. The other is based on the 1990 tabulation block information and will be referred to as the estimated HU count. This second count will be used later by the Long Form Sample Design and Estimation Team.

d. Append the estimated 1990 urbanicity of each 2000 collection block to the Universe File. This information can be found on the Merged Data File.

e. Sort the Universe File by county, block cluster number, and 2000 collection block.

f. Sum each of the 12 demographic/tenure counts, the two tenure counts, and the estimated HU and occupied HU counts across all 2000 collection blocks in a 2000 collection block cluster. On the Universe File, append these 16 block cluster totals to each collection block in the block cluster.

g. For each 2000 collection block cluster, examine the current HU count (from block clustering) for all collection blocks in the cluster. Choose the estimated 1990 urbanicity of the 2000 collection block with the most HUs. Assign that value to the estimated urbanicity of the 2000 collection block cluster for each collection block in the cluster.

If two or more blocks tie for the most HUs then examine the estimated 1990 urbanicity of all those that tie. Choose the lowest value (closest to 0) among them. Assign that value to the estimated 1990 urbanicity of the 2000 collection block cluster for each collection block in the cluster.

2. Create Sampling and Sorting Variables

Now we will create some variables that will be used during the sampling and sorting operations. Assign the cluster variable created below to all 2000 collection blocks in the cluster.

a. Since different sampling operations will be performed on different size block clusters, we want to classify clusters into three size categories -- small, medium, and large. Create a variable for cluster size as follows.

Table 3. Block Cluster Size Classification Rules

IF	THEN
HU count (From block clustering)	Size Category
0-2	1
3-79	2
80+	3

b. For 24 states, DC, and Puerto Rico there will be three sampling strata. These are the areas with few or no American Indians on reservation. For the other 26 states that have an American Indian population on reservation there will be four sampling strata. Use the number of sampling strata variable on the Sample Size Input File (ACE2000_TB.FIN). See Attachment K for a file layout. Attachment G also has a listing of which states fall into which category.

For each 2000 collection block, create a variable which record the number of sampling strata in the state. In states with three strata (those without American Indian Reservation sample), the strata will be defined as follows:

Table 4. Sampling Strata: States without American Indian Reservation Sample

IF	THEN
Size Category	Sampling stratum
1	1
2	2
3	3

In states with four strata (those with AIR sample), small block clusters on Reservations will be grouped with other small block clusters. Medium and large block clusters on Reservations will be placed in the same strata. Documentation for the AIR sample is forthcoming.

Table 5. Sampling Strata: States with American Indian Reservation Sample

IF		THEN
Size Category	American Indian Country Indicator	Sampling stratum
1	0, 1, or 2	1
2	0 or 2	2
2	1	4
3	0 or 2	3
3	1	4

c. To help make sure we have a balanced representation of all groups in our sample we are going to create a sort variable based on the demographic/tenure make-up of each block cluster. Create this demographic/tenure variable and a label as follows.

i. For each 2000 block cluster, divide the number of people in each of the 12 demographic/tenure groups by the total number of people to determine the population proportion of each group. Use the same method to also determine the population proportion of renters and owners in the block cluster. Round each value to three decimal places (x.0005 rounds to x.001). Append these values to the Universe File.

ii. For the 50 states and DC, assign the cluster to a demographic/tenure group based on the following ordered rules. These rules are based on research that will be documented at a later date.

The first rule that a cluster satisfies is the group to which the cluster is assigned. Once a cluster is assigned, stop processing that cluster.

Table 6. Demographic/Tenure Group Assignment Rules for the 50 States and the District of Columbia

	IF	THEN	
Order	Criteria	Dem/Ten Group Code	Dem/Ten Group Label
1	Proportion of Hawaiian and Pacific Islander Renters ≥ 0.10	1	PR
2	Proportion of Hawaiian and Pacific Islander Owners ≥ 0.10	2	PO
3	Proportion of American Indian and Alaska Native Renters ≥ 0.10	5*	IR
4	Proportion of American Indian and Alaska Native Owners ≥ 0.10	6*	IO
5	Proportion of Asian Renters ≥ 0.20	3*	AR
6	Proportion of Asian Owners ≥ 0.20	4*	AO
7	Proportion of Hispanic Renters ≥ 0.20	7	HR
8	Proportion of Hispanic Owners ≥ 0.20	8	HO
9	Proportion of Black Renters ≥ 0.25	9	BR
10	Proportion of Black Owners ≥ 0.25	10	BO
11	Proportion of Other Renters ≥ 0.30	11	OR
12	all else	12	OO

*Smaller demographic groups are given precedence in the assignment process. Within each race group, renter is also given precedence over owner. This order does not coincide with the demographic/tenure group codes, which are assigned so that certain groups will be sorted together before selection.

iii. For Puerto Rico, assign the cluster to a demographic/tenure group based on the following rules. The first rule that a cluster satisfies is the group the cluster is assigned to. [Note: The groups are listed in ascending order based on the number of members.]

Table 7. Demographic/Tenure Group Assignment Rules for Puerto Rico

	IF	THEN	
Order	Criteria	Dem/Ten Group Code	Dem/Ten Group Label
1	Proportion of Renters ≥ 0.45	13	TR
2	all else	14	TO

d. Assign each collection block a region code and a division code. This information is contained in the Sample Size Input File and in Attachment G. These are based on the standard definitions used by the GEO.

3. Produce Universe File Verification Outputs

a. We will evaluate how the Demographic/Tenure Group variable classifies the block clusters. As such, for each state, DC and Puerto Rico, tally the demographic/tenure counts and the HU and block cluster counts in a sampling strata for each Demographic/Tenure Group compute several person and HU counts. Write this information to a file. See Attachment H for a file layout and example.

b. We would also like to evaluate the estimated 1990 urbanicity variable. For each state, calculate the proportion of 2000 collection blocks in each estimated 1990 urbanicity group. Then calculate the number of people that live in blocks with each of the three classifications. Also, calculate the proportion of 2000 collection block clusters that fall into each urbanicity group and the number of people that live in block clusters which fall into each urbanicity group. Write this information to a file. See Attachment I for a layout.

c. Provide the DSSD access to the Universe File for verification. See Attachment J for a file layout of the Universe File.

IV. Creation of the Block Cluster Sampling Parameter File

The parameters for selecting the initial sample of block clusters are specified in this section. These parameters will be an input to a forthcoming memorandum specifying the selection of the block cluster sample.

The block cluster sampling will be done separately for each of the four sampling strata within each state. As part of the Universe File creation in Section III, all clusters were assigned to one of four sampling strata: 1) Small block clusters, 2) Medium block clusters, 3) Large block clusters, and 4) AIR block clusters[2]. Each of these four categories will be sampled at different rates during the initial block cluster sampling.

Approximately 25,000 block clusters were originally allocated to the 50 states and the District of Columbia[3]. This allocation was recently updated, resulting in the target number of medium and large sample block clusters for each state as shown in Attachment G. Attachment G also contains the target number of small block clusters for each state, and the target number of clusters for Puerto Rico. These numbers, which are needed for the calculations that follow, can be obtained from the Sample Size Input File, which is described in the Input Section below. This file will be the basis for the Block Cluster Sampling Parameter File. The procedures for calculating the first-step sampling parameters for each state are below. Do these steps separately for each sampling stratum.

A. Calculation of the Sampling Parameters

The calculation of the sampling parameters requires two basic pieces of information for each state and sampling stratum: the sample size and the universe count. In general, the take-every is the ratio of the universe count to the sample size. The calculation of the large block sampling stratum take-every has one additional calculation which controls the sample so that the expected listing does not exceed the budgeted listing.

1. Sample Sizes

Obtain the following information for the state from the Sample Size Input File (see Attachment K for a layout):

---

[2]Not all states will have block clusters classified into sampling stratum 4. See Attachment G.

[3]Schindler, E. (1998), "Allocation of the ICM Sample to the States for Census 2000," Proceedings of the Section on Survey Research, American Statistical Association.

Budgeted Listing for medium and large clusters, BLIST
Target number of clusters for small sampling stratum, TCLUSTS
Target number of clusters for the medium and large sampling strata, TCLUST
Target number of clusters in AIR stratum, TCLUSTA

2. Universe File Counts

For each state use the Universe File to produce the total number of block clusters and the total number of HUs for each of the four sampling strata. The counts are:

Total number of housing units in small clusters, NHUS
Total number of small block clusters, NCLUSTS
Total number of housing units in medium clusters, NHUM
Total number of medium clusters, NCLUSTM
Total number of housing units in large clusters, NHUL
Total number of large clusters, NCLUSTL
Total number of housing units in AIR, NHUA
Total number block clusters in AIR, NCLUSTA.

3. Allocate Block Clusters to Sampling Strata

For the small block cluster and AIR sampling strata, the target clusters are given on the Sample Size Input File. For the medium and large sampling strata, a combined target has been provided on the Sample Size Input File. This combined target needs to be proportionally allocated to the large and medium sampling strata.

Calculate the target number of medium block clusters and the target number of large block clusters by proportionally allocating the combined target to the medium and large sampling strata based on HUs.

$$\text{TCLUSTM} = \frac{\text{NHUM}}{\text{NHUM} + \text{NHUL}} \times \text{TCLUST}$$

$$\text{TCLUSTL} = \frac{\text{NHUL}}{\text{NHUM} + \text{NHUL}} \times \text{TCLUST}$$

4. Check if Listing is Within Budget

A listing adjustment is applied to the large block cluster take-every to aid in controlling the listing workload. If the preliminary expected number of

HUs is larger than the budgeted listing, an adjustment needs to be made to the large block sampling rate to reduce the expected number of HUs.

PRELIST is the preliminary number of expected HUs to be listed from the medium and large sampling strata.

$$PRELIST = TCLUSTM \times \frac{NHUM}{NCLUSTM} + TCLUSTL \times \frac{NHUL}{NCLUSTL}$$

Compare the preliminary expected listing to the budgeted listing and calculate the listing adjustment, LISTADJ. Round LISTADJ to 6 decimal places.

If BLIST < PRELIST, then

$$LISTADJ = \frac{BLIST - TCLUSTM \times \frac{NHUM}{NCLUSTM}}{TCLUSTL \times \frac{NHUL}{NCLUSTL}}$$

If BLIST ≥ PRELIST, then LISTADJ = 1.000000

5. Calculate Sampling Parameters

Calculate the take-everys and random starts for the four sampling strata. The take-everys are calculated by dividing the total number of clusters by the target number of clusters. Then if necessary, the large block cluster take-every is then adjusted to comply with the budgeted listing. Random starts are calculated by multiplying the take-every by a random number between zero and one ($0 < RN \leq 1$). For each state and sampling stratum (SS=1, ..., 4) obtain the appropriate number of clusters from step two and the corresponding target from step three. Generate a new random number for each of the four sampling strata.

Calculate the take-every:

For sampling strata 1, 2 and 4:

$$TE1_{SS} = \frac{NCLUST_{SS}}{TCLUST_{SS}}$$ (Round to 6 decimal places.)

For sampling stratum 3:

$$TE1_{SS} = \frac{NCLUST_{SS}}{TCLUST_{SS} \times LISTADJ} \quad \text{(Round to 6 decimal places.)}$$

Calculate a random start for each of the four sampling strata:

$$RS1_{SS} = RN \times TE1_{SS}. \quad \text{(Round to 6 decimal places.)}$$

B. Calculation of Expected Number of Housing Units and Clusters for Listing

These expected values will be placed on the Sample Summary File and compared to the sampling results. For each state and sampling strata obtain the total number of clusters from step two and the take-every from step five in section IV.A. Calculate the expected number of clusters and HUs for each sampling stratum:

$$ECLUSTL_{SS} = \frac{NCLUST_{SS}}{TE1_{SS}} \quad \text{(Round to the nearest integer.)}$$

$$EXPHUL_{SS} = \frac{NHU_{SS}}{TE1_{SS}} \quad \text{(Round to the nearest integer.)}$$

C. Creation of the Sample Summary File and Block Cluster Sampling Parameter File

1. Sample Summary File

The Sample Summary File is a state level file and is one of the two files used to provide a history of the ACE sample[4]. A description and layout of the file was documented in a previous memorandum, but the initial version will be created in this section. Future versions of the file will be created as updates are made following each sampling operation. Using the information calculated for each state and sampling stratum in section IV.A, create version one of the Sample Summary File, ACE2000_SSFV1.<mmddyy>. The extension <mmddyy> is the date the file is created (i.e. 123199 is the extension for a file created on December 31, 1999). Include the following variables:

1. Census Region, REGION
2. Census Division, DIV

[4]The second file used in tracking the ACE sample is the Sample Design File will be created following the initial block cluster sampling. Descriptions of these files are documented in the memorandum from Kostanich to Stoudt, "Accuracy and Coverage Evaluation (ACE) Survey: Documentation for the Sample Summary File and Sample Design File," March 1999.

3. State code (01-72 = FIPS State Code), STATE
4. Budgeted Listing for medium and large clusters, BLIST
5. Target number of medium and large block clusters, TCLUST
6. Target number of small block clusters, TCLUSTS
7. Target number of AIR block clusters, TCLUSTA
8. Total number of block clusters, NCLUST

$$NCLUST = \sum_{ss=1}^{4} NCLUST_{ss}$$

9. Total number of housing units, NHU

$$NHU = \sum_{ss=1}^{4} NHU_{ss}$$

10. Expected clusters in sample to list, ECLUSTL

$$ECLUSTL = \sum_{ss=1}^{4} ECLUSTL_{ss}$$

11. Expected housing units in sample to list, EXPHUL

$$EXPHUL = \sum_{ss=1}^{4} EXPHUL_{ss}$$

2. The Block Cluster Sampling Parameter File

The Block Cluster Sampling Parameter File contains one record for each sampling stratum in each of the 50 states, the District of Columbia, and Puerto Rico. During production, a separate file will be created for each state, the District of Columbia and Puerto Rico. These files will serve as inputs to the block cluster sample selection specification. Following the completion of the sample selection for all states, the state files will be concatenated into one file. A description of this file, ACE2000_PARABC.<mmddyy>, is provided in section V. Using the information from section IV.A for each state and sampling stratum, create the Block Cluster Sampling Parameter Files. Include the following variables:

1. Census Region, REGION
2. Census Division, DIV
3. State code (01-72 = FIPS State Code), STATE
4. Sampling Stratum, SS
5. Target number of block clusters, TCLUST
6. Total number of block clusters, NCLUST
7. Total number of housing units, NHU
8. First-step take-every, TE1
9. First-step random start, RS1

D. Block Cluster Sampling Parameter File Verification Output

Using the Block Cluster Sampling Parameter File and a listing of the random numbers generated to calculate the random starts, the DSSD will conduct an independent verification of the parameter calculations.

V. Input

The following files will be needed for producing both the Universe File and the Block Cluster Sampling Parameter File.

A. Block Cluster Files

These are block level files which contain information about which blocks are clustered together. There will be one file for each state. These files are created by the GEO as specified in the memorandum from Hogan to Marx, "Census 2000 Specifications for Block Cluster Formation," February 16, 1999. See Attachment B for a file layout.

B. 1990 Hundred Percent Edited Detail File

This file contains data collected from the 1990 census. We will get both person information and 1990 tabulation block information from this file. This is the source of person-level demographic and tenure information.

C. Block Equivalency Files

These files show the relationships between the 1990 tabulation blocks and the 2000 collection blocks. In these files there is a record for each 1990 tabulation block/2000 collection block relationship. There is one file per county. See Attachment A for a file layout.

D. 1990 Geographic Reference File

This file contains geographic and legal information about 1990 tabulation and collection blocks.

E. Sample Size Input File: ACE2000_TB.FIN

This file contains information about each state that is needed for the sampling process. This file will be provided by the DSSD. There is one record for each state, the District of Columbia, and Puerto Rico. See Attachment K for a file layout.

Attachment A

Block Equivalency File Layout

Variable Description	Name	Places
1990 Tabulation State	STATE90	1-2
1990 Tabulation County	COUNTY90	4-6
1990 Tabulation Census Tract	TRACT90	8-11
1990 Tabulation Tract Suffix	TRASUF90	12-14
1990 Tabulation Block	TBLOCK90	15-17
1990 Tabulation Suffix	TBSUF90	18-18
2000 Collection State	STATE2K	20-21
2000 Collection County	COUNTY2K	23-25
2000 Collection Block	CBLOCK2K	27-31

Block Cluster File Layout

Variable Description	Name	Places
State	STATE	1-2
County	COUNTY	3-5
Interim Tract (a.k.a. pseudo-tract)	ITRACT	6-11
Block Number	COLBLOCK	12-16
Blank		17-17
Cluster Number (geography not ACE)	GCLUS	18-22
Blank		23-23
Cluster Size code	CLUSSIZE	24-24
1 = Clusters with 0 HUs		
2 = Clusters with 1 HUs		
3 = Clusters with 2 HUs		
4 = Clusters with between 3 and 5 HUs		
5 = Clusters with between 6 and 9 HUs		
6 = Clusters with between 10 and 19 HUs		
7 = Clusters with between 20 and 29 HUs		
8 = Clusters with between 30 and 79 HUs		
9 = Clusters with 80 or more HUs		
Blank		25-25
Block Area (Sq. Miles)	BAREA	26-33
Blank		34-34
Block Perimeter (Miles)	BPERIM	35-40
Blank		41-41
Block Cluster Area (Sq. Miles)	BCAREA	42-49
Blank		50-50
Block Cluster Perimeter (Miles)	BCPERIM	51-56
Number of HUs in cluster	NHU	57-61
Number of HUs in block	NHUBLOCK	62-66
Block TEA	TEA	67-67
1 = Mailout/Mailback		
2 = Update/Leave		
3 = List/Enumerate		
5 = Rural Update/Enumerate		
6 = Military		
7 = Urban Update/Leave		
8 = Update/Leave to Mailout/Mailback conversions		
9 = Mailout/Mailback to Update/Leave conversions		

TEA Group for Block Cluster A= Mailout/Mailback or Urban Update/Leave or Update/Leave to Mailout/Mailback conversions B= Update/Leave or Rural Update/Enumerate C=List/Enumerate D=Military E=Mailout/Mailback to Update/Leave conversions	TEABC	68-68
2000 MAF HUs count ' ' Blank if no HU count available	NHUM	69-73
1990 ACF HUs count ' ' Blank if no HUs count available	NHU90	74-78
Housing Unit Count Indicator 1 = from 2000 MAF 2 = from 1990 ACF	HUIND	79-79
Invisible Boundary Collapse Indicator 0 = No 1 = Yes (Collapsing across Invisible Boundary in BC)	INV	80-80
American Indian Country Indicator 0 = No American Indian Country 1 = American Indian Reservation/trust land 2 = Tribal jurisdiction statistical area/ Alaska Native Village statistical area/ tribal designated statistical area	AICIND	81-81
Military Indicator 0 = No Military Area 1 = Block contains Military Area	MILIND	82-82
Collapsed Enclosed Block Indicator 0 = Otherwise 1 = An enclosed block has been forced to collapse	CEBI	83-83

Definitions of the Demographic and Tenure Characteristics

Using variables Q4 (race), Q7 (Hispanic origin), and H4 (tenure) from the 1990 HEDF we can assign any person to one of the 12 demographic/tenure groups.

Questions Q4 and Q7 are used to determine each person's demographic group. Question Q4 asks for the respondent's race. Write-in responses are coded to the values 001 to 946. Values 971 to 986 are used for marked responses which include white, black, American Indian, Eskimo, Aleut, Chinese, etc. Question Q7 asks for a person's Hispanic origin. Values 2-5 indicate different Hispanic origins (Cuban, Puerto Rican, etc.)

The six demographic categories have been designed to be mutually exclusive. So, if a person is of Hispanic Origin that person would be classified as Hispanic regardless of race. The following algorithm has been developed to classify each person.

American Indian and Alaska Native =	([Q4 ≥ 000 AND Q4 ≤ 599] OR [Q4 ≥ 935 AND Q4 ≤ 940] OR [Q4 ≥ 941 AND Q4 ≤ 970] OR [Q4 ≥ 973 AND Q4 ≤ 975]) AND Q7≤ 1
Asian =	([Q4 ≥ 600 AND Q4 ≤ 652] OR [Q4 ≥ 976 AND Q4 ≤ 977] OR [Q4 ≥ 979 AND Q4 ≤ 982] OR Q4 = 985) AND Q7≤ 1
Black =	([Q4 ≥ 870 AND Q4 ≤ 934] OR Q4 = 972) AND Q7≤ 1
Hawaiian and Pacific Islander =	([Q4 ≥ 653 AND Q4 ≤ 699] OR Q4 = 978 OR Q4 = 983 OR Q4 = 984) AND Q7≤ 1
Hispanic =	Q7 ≥ 2
White and Other =	Remaining records (any records left over after all other race/ethnic group criteria are exhausted)

Tenure is defined by the H4 variable as follows:

Owner = H4 = 1 OR H4 = 2

Renter = H4 = 3 OR H4 = 4

To determine the number of HUs and the number of occupied HUs use the variable HC1. The variable HC1 has the following values:

0 = occupied
1 = for rent
2 = for sale only
3 = rented or sold, not occupied
4 = for seasonal/recreational/occasional use
5 = for migratory workers
6 = other vacant

Occupied = HC1 = 0
Vacant = HC1 $\neq$ 0

Attachment D

Size of Urban Area

The following are the possible values for the urban area size of a 1990 tabulation block.

Code	Number of People
0	not in universe
1	0
2	1 - 24
3	25 - 99
4	100 - 199
5	200 - 249
6	250 - 299
7	300 - 499
8	500 - 999
9	1,000 - 1,499
10	1,500 - 1,999
11	2,000 - 2,499
12	2,500 - 4,999
13	5,000 - 9,999
14	10,000 - 19,999
15	20,000 - 24,999
16	25,000 - 49,999
17	50,000 - 99,999
18	100,000 - 249,000
19	250,000 - 499,999
20	500,000 - 999,999
21	1,000,000 - 2,499,999
22	2,500,000 - 4,999,999
23	5,000,000 or more

Merged Data File: Verification Counties

For verification of the Merged Data File, provide the DSSD with the files for counties listed below. These are all from wave one states.

Bear Lake County, Idaho (FIPS = 16007)
Grand Forks County, North Dakota (FIPS = 38035)
Anchorage Borough, Alaska (FIPS = 02020)
Rosebud County, Montana (FIPS = 30087)
Blue Earth County, Minnesota (FIPS = 27013)
District of Columbia (the entire district is considered a county) (FIPS = 11001)
Shannon County, South Dakota (FIPS = 46113)
Menominee County, Wisconsin (FIPS = 55078)

Merged Data File Layout

Variable Description	Name	Places
1990 Tabulation State	STATE90	1-2
1990 Tabulation County	COUNTY90	4-6
1990 Tabulation Census Tract	TRACT90	8-11
1990 Tabulation Tract Suffix	TRASUF90	12-14
1990 Tabulation Block	TBLOCK90	15-17
1990 Tabulation Suffix	TBSUF90	18-18
2000 Collection State	STATE2K	20-21
2000 Collection County	COUNTY2K	23-25
2000 Collection Block	CBLOCK2K	27-31
Blank		32-35
2000 Collection Block Housing Unit Count	CB2KHU	36-40
Geography Block Cluster Number	GCLUST	41-45
Blank		46-49
1990 Tabulation Block Number of:		
Hawaiian and Pacific Islander Renter	PIR90TAB	50-54
Hawaiian and Pacific Islander Owner	PIO90TAB	55-59
American Indian and Alaska Native Renter	IR90TAB	60-64
American Indian and Alaska Native Owner	IO90TAB	65-69
Asian Renter	AR90TAB	70-74
Asian Owner	AO90TAB	75-79
Hispanic Renter	HR90TAB	80-84
Hispanic Owner	HO90TAB	85-89
Black Renter	BR90TAB	90-94
Black Owner	BO90TAB	95-100
White and Other Renter	OR90TAB	100-104
White and Other Owner	OO90TAB	105-109
Total Renters	R90TAB	110-114
Total Owners	O90TAB	115-119
Total Housing Units	HU90TAB	120-124
Occupied Housing Units	OHU90TAB	125-129
Total People (Non-GQ)	POP90TAB	130-134
Blank		135-145
1990 Urban/Rural (UR)	UR90	146-146
1990 Urban Area Size (UASZ)	UASZ90	147-148
# of 2000 collection block HUs associated with 1990 tabulation block	HU2KA90	149-153
Proportion of HUs in 2000 collection block associated with 1990 tabulation block	P	154-161

Blank		162-170
1990 Tab/2000 Collection Association Estimated Number of:		
Hawaiian and Pacific Islander Renter	PIRASSOC	171-175
Hawaiian and Pacific Islander Owner	PIOASSOC	176-180
American Indian and Alaska Native Renter	IRASSOC	181-185
American Indian and Alaska Native Owner	IOASSOC	186-190
Asian Renter	ARASSOC	191-195
Asian Owner	AOASSOC	196-200
Hispanic Renter	HRASSOC	201-205
Hispanic Owner	HOASSOC	206-210
Black Renter	BRASSOC	211-215
Black Owner	BOASSOC	216-220
White and Other Renter	ORASSOC	221-225
White and Other Owner	OOASSOC	226-230
Total Renters	RASSOC	231-235
Total Owners	OASSOC	236-240
Total Housing Units	HUASSOC	241-245
Occupied Housing Units	OHUASSOC	246-250
Blank		251-260
2000 Collection Block Estimated Number of:		
Hawaiian and Pacific Islander Renter	ECOLPIR	261-265
Hawaiian and Pacific Islander Owner	ECOLPIO	266-270
American Indian and Alaska Native Renter	ECOLIR	271-275
American Indian and Alaska Native Owner	ECOLIO	276-280
Asian Renter	ECOLAR	281-285
Asian Owner	ECOLAO	286-290
Hispanic Renter	ECOLHR	291-295
Hispanic Owner	ECOLHO	296-300
Black Renter	ECOLBR	301-305
Black Owner	ECOLBO	306-310
White and Other Renter	ECOLOR	311-315
White and Other Owner	ECOLOO	316-320
Total Renters	ECOLR	321-325
Total Owners	ECOLO	326-330
Total Housing Units	ECOLHU	331-335
Occupied Housing Units	ECOLOHU	336-340
Blank		341-350
Estimated total population of 2000 collection block	ECOLPOP	351-355
1990 tabulation block urbanicity	90TABURB	356-356
Estimated 1990 urbanicity of 2000 collection block	ECOLURB	357-357

Attachment G

Sampling Information for Each State

Region	Division	State	Number of Strata	Budgeted Listing	AIR Cluster Target	Small Cluster Target	Medium and Large Cluster Target	First Block Cluster Number
3	6	Alabama*	3	25,347	0	116	417	41001
4	9	Alaska	4	27,196	1	20	334	91001
4	8	Arizona	4	48,451	110	86	492	81001
3	7	Arkansas	3	24,744	0	90	494	51001
4	9	California	4	284,076	14	184	2,753	92001
4	8	Colorado	4	37,965	2	83	479	82001
1	1	Connecticut*	3	30,039	0	20	377	11001
3	5	Delaware	3	21,610	0	20	413	31001
3	5	DC	3	53,369	0	20	384	32001
3	5	Florida	4	62,845	1	145	520	33001
3	5	Georgia*	3	37,384	0	154	399	34001
4	9	Hawaii	3	45,059	0	20	300	97001
4	8	Idaho	4	19,157	6	54	412	83001
2	3	Illinois	3	31,571	0	185	430	61001
2	3	Indiana	3	15,925	0	140	275	62001
2	4	Iowa*	3	14,108	0	147	300	71001
2	4	Kansas	4	16,281	1	193	300	72001
3	6	Kentucky	3	29,621	0	96	447	42001
3	7	Louisiana*	3	37,378	0	65	595	52001
1	1	Maine	4	16,572	1	38	309	12001
3	5	Maryland	3	41,107	0	36	368	35001
1	1	Massachusetts*	3	27,255	0	38	375	13001
2	3	Michigan	4	24,128	5	122	379	63001
2	4	Minnesota	4	19,091	10	141	300	73001
3	6	Mississippi	4	19,990	3	81	402	43001
2	4	Missouri	3	19,807	0	162	300	74001
4	8	Montana	4	17,969	24	67	420	84001
2	4	Nebraska	4	14,177	3	142	300	75001
4	8	Nevada	4	63,031	5	46	468	85001
1	1	New Hampshire	3	21,128	0	25	307	14001
1	2	New Jersey**	3	37,394	0	39	461	21001
4	8	New Mexico	4	32,242	70	108	481	86001
1	2	New York	4	143,949	5	143	1,261	22001
3	5	North Carolina	4	26,717	4	143	400	36001
2	4	North Dakota	4	15,738	12	121	300	76001
2	3	Ohio	3	30,790	0	132	421	64001
3	7	Oklahoma	4	25,328	8	142	426	53001
4	9	Oregon	4	20,577	3	86	320	98001

Demographic/Tenure Group Evaluation File Layout

Variable Description	Name	Places
State	STATE	1-2
Demographic/Tenure group code (for Puerto Rico only include codes 13-14, for all others only include codes 1-12)	DTCODE	4-5
Demographic/Tenure group label	DTLABEL	7-8
For each of the following counts record the number in this demographic group in this state		
# of small block clusters	SMCLUS	10-17
# of medium block clusters	MEDCLUS	18-25
# of large block cluster	LARCLUS	26-33
# of housing units in small block clusters	HUSMC	34-41
# of housing units in medium block cluster	HUMC	42-49
# of housing units in large block cluster	HULC	50-57
Est. # of Hawaiian and Pacific Islander Renters	ESTPR	58-65
Est. # of Hawaiian and Pacific Islander Owners	ESTPO	66-73
Est. # of Asian Renters	ESTAR	74-81
Est. # of Asian Owners	ESTAO	82-89
Est. # of Am. Indian Renters	ESTIR	90-97
Est. # of Am. Indian Owners	ESTIO	98-105
Est. # of Hispanic Renters	ESTHR	106-113
Est. # of Hispanic Owners	ESTHO	114-121
Est. # of Black Renters	ESTBR	122-129
Est. # of Black Owners	ESTBO	130-137
Est. # of White and Other Renters	ESTOR	138-145
Est. # of White and Other Owners	ESTOO	146-153
Est. # of Renters	ESTR	154-161
Est. # of Owners	ESTO	162-169
Estimated total population	ESTPOP	170-178

The following is an example of how the file will look.

State	Demographic/Tenure Group Code	Demographic/Tenure Group Label	Small Clusters	.... # of HUs in Large Cluster	PR	PO	AR..
AL	1	PR	5	548	1,000	578	45..
AL	2	PO	7	0	456	984	74..
AL	3	AR					
AL . . .							
.							
AK	1	PR . . .					
AK	2	PO . . .					
.							
AR	1.....						

Attachment I

Urbanicity Verification File Layout

In order to verify the assignment of the urbanicity variable we would like a file with the following layout. If after either of the transitions, from 1990 tabulation to 2000 collection or from 2000 collection to 2000 cluster, there are great differences in the proportions then we will know that we should examine the transition further. The urbanicity values will also be verified on a micro-level in each of the selected counties.

Variable Description	Name	Places
State	STATE	1-2
Proportion of 2000 collection blocks with urbanicity = 1	PCBU1	3-7
Proportion of 2000 collection blocks with urbanicity = 2	PCUB2	8-12
Proportion of 2000 collection blocks with urbanicity = 3	PCUB3	13-17
People who live in 2000 collection blocks with urbanicity = 1	PEPCUB1	18-28
People who live in 2000 collection blocks with urbanicity = 2	PEPCUB2	29-39
People who live in 2000 collection blocks with urbanicity = 3	PEPCUB3	40-50
Blank		51-59
Proportion of 2000 collection blocks clusters with urbanicity = 1	PCLUU1	60-64
Proportion of 2000 collection blocks clusters with urbanicity = 2	PCLUU2	65-69
Proportion of 2000 collection blocks clusters with urbanicity = 3	PCLUU3	70-74
People who live in 2000 collection blocks clusters with urbanicity = 1	PEPCLUU1	75-85
People who live in 2000 collection blocks clusters with urbanicity = 2	PEPCLUU2	86-96
People who live in 2000 collection blocks clusters with urbanicity = 3	PEPCLUU3	97-107

Universe File Layout

Variable Description	Name	Places
State	STATE	1-2
County	COUNTY	3-5
Interim Tract (a.k.a. pseudo-tract)	ITRACT	6-11
Block Number	COLBLOCK	12-16
Blank		17-17
Cluster Number (geography not ACE)	GCLUS	18-22
Blank		23-23
Cluster Size code 1 = Clusters with 0 HUs 2 = Clusters with 1 HUs 3 = Clusters with 2 HUs 4 = Clusters with between 3 and 5 HUs 5 = Clusters with between 6 and 9 HUs 6 = Clusters with between 10 and 19 HUs 7 = Clusters with between 20 and 29 HUs 8 = Clusters with between 30 and 79 HUs 9 = Clusters with 80 or more HUs	CLUSSIZE	24-24
Blank		25-25
Block Area (Sq. Miles)	BAREA	26-33
Blank		34-34
Block Perimeter (Miles)	BPERIM	35-40
Blank		41-41
Block Cluster Area (Sq. Miles)	BCAREA	42-49
Blank		50-50
Block Cluster Perimeter (Miles)	BCPERIM	51-56
Number of HUs in cluster	NHU	57-61
Number of HUs in block	NHUBLOCK	62-66
Block TEA 1 = Mailout/Mailback 2 = Update/Leave 3 = List/Enumerate 5 = Rural Update/Enumerate 6 = Military 7 = Urban Update/Leave 8 = Update/Leave to Mailout/Mailback conversions 9 = Mailout/Mailback to Update/Leave conversions	TEA	67-67

TEA Group for Block Cluster A= Mailout/Mailback or Urban Update/Leave or Update/Leave to Mailout/Mailback conversions B= Update/Leave or Rural Update/Enumerate C=List/Enumerate D=Military E=Mailout/Mailback to Update/Leave conversions	TEABC	68-68
2000 MAF HUs count ' ' Blank if no HU count available	NHUM	69-73
1990 ACF HUs count ' ' Blank if no HUs count available	NHU90	74-78
Housing Unit Count Indicator 1 = from 2000 MAF 2 = from 1990 ACF	HUIND	79-79
Invisible Boundary Collapse Indicator 0 = No 1 = Yes (Collapsing across Invisible Boundary in BC)	INV	80-80
American Indian Country Indicator 0 = No American Indian Country 1 = American Indian Reservation/trust land 2 = Tribal jurisdiction statistical area/ Alaska Native Village statistical area/ tribal designated statistical area	AICIND	81-81
Military Indicator 0 = No Military Area 1 = Block contains Military Area	MILIND	82-82
Collapsed Enclosed Block Indicator 0 = Otherwise 1 = An enclosed block has been forced to collapse	CEBI	83-83

Blank		84-90
2000 Collection Block Estimated Number of:		
Hawaiian and Pacific Islander Renter	ECOLPIR0	91-95
Hawaiian and Pacific Islander Owner	ECOLPIO	96-100
American Indian and Alaska Native Renter	ECOLIR	101-105
American Indian and Alaska Native Owner	ECOLIO	106-110
Asian Renter	ECOLAR	111-115
Asian Owner	ECOLAO	116-120
Hispanic Renter	ECOLHR	121-125
Hispanic Owner	ECOLHO	126-130
Black Renter	ECOLBR	131-135
Black Owner	ECOLBO	136-140
White and Other Renter	ECOLOR	141-145
White and Other Owner	ECOLOO	146-150
Total Renters	ECOLR	151-155
Total Owners	ECOLO	156-160
Total Housing Units	ECOLHU	161-165
Occupied Housing Units	ECOLOHU	166-170
Total People (Non-GQ)	ECOLPOP	171-175
Estimated 1990 urbanicity of the 2000 collection block	ECOLURB	176-176
1 = Urban Area with 1990 population ≥ 250,000		
2 = Other Urban Area		
3 = Non-Urban Area		
Blank		177-180
2000 Collection Block Cluster Estimated Number of:		
Hawaiian and Pacific Islander Renter	ECLUSPIR	181-185
Hawaiian and Pacific Islander Owner	ECLUSPIO	186-190
American Indian and Alaska Native Renter	ECLUSIR	191-195
American Indian and Alaska Native Owner	ECLUSIO	196-200
Asian Renter	ECLUSAR	201-205
Asian Owner	ECLUSAO	206-210
Hispanic Renter	ECLUSHR	211-215
Hispanic Owner	ECLUSHO	216-220
Black Renter	ECLUSBR	221-225
Black Owner	ECLUSBO	226-230
White and Other Renter	ECLUSOR	231-235
White and Other Owner	ECLUSOO	236-240
Total Renters	ECLUSR	241-245
Total Owners	ECLUSO	246-250
Total Housing Units	ECLUSHU	251-255
Occupied Housing Units	ECLUSOHU	256-260
Total People (Non-GQ)	ECLUSPOP	261-265
Blank		266-275

Estimated 1990 urbanicity of 2000 block cluster	ECLUSURB	276-276
1 = Urban Area with 1990 population ≥ 250,000		
2 = Other Urban Area		
3 = Non-Urban Area		
Size Category	SIZECAT	277-277
1 = Small (0-2 HUs)		
2 = Medium (3-79 HUs)		
3 = Large (80+ HUs)		
Number of sampling strata in state	NSSINST	278-278
Sample stratum	SS	279-279
1 = Small		
2 = Medium (non-AIR)		
3 = Large (non-AIR)		
4 = American Indian Reservation		
Blank		280-285
2000 Collection Block Cluster Proportion of Population that is:		
Hawaiian and Pacific Islander Renter	CLUPPIR	286-290
Hawaiian and Pacific Islander Owner	CLUPPIO	291-295
American Indian and Alaska Native Renter	CLUPIR	296-300
American Indian and Alaska Native Owner	CLUPIO	301-305
Asian Renter	CLUPAR	306-310
Asian Owner	CLUPAO	311-315
Hispanic Renter	CLUPHR	316-320
Hispanic Owner	CLUPHO	321-325
Black Renter	CLUPBR	326-330
Black Owner	CLUPBO	331-335
White and Other Renter	CLUPOR	336-340
White and Other Owner	CLUPOO	341-345
Renters	CLUPR	346-350
Owners	CLUPO	351-355
Blank		356-364
Demographic/Tenure group (code)	DTCODE	365-366
Demographic/Tenure group (label)	DTLABEL	367-368
Region	REGION	369-369
Division	DIV	370-370

UNITED STATES DEPARTMENT OF COMMERCE
Bureau of the Census
Washington, DC 20233

MAY 03 1999

DSSD CENSUS 2000 PROCEDURES AND OPERATIONS MEMORANDUM SERIES # R-7

MEMORANDUM FOR Howard Hogan
Chief, Decennial Statistical Studies Division

From: Donna Kostanich
Assistant Division Chief, Sampling and Estimation
Decennial Statistical Studies Division

Subject: Accuracy and Coverage Evaluation Survey: Sample Size Estimates

This memorandum documents our current estimates of housing unit sample size for the Accuracy and Coverage Evaluation (ACE) survey for budgeting purposes. These sample sizes could change due to the budget process or operational resource constraints. Furthermore, the actual sample sizes may be different to the extent the universe differs from the assumptions we used to estimate these numbers.

These numbers are presented for the total U.S. which includes the 50 states and the District of Columbia, and are consistent with the approximately 300,000 housing unit ACE design. Furthermore, these numbers assume that the ACE sample design is contingent on the Integrated Coverage Measurement (ICM) sample design which is the basis of the listing and block cluster numbers.

The housing unit sample sizes in terms of listing and interviewing are provided in the table. These numbers are approximately 1,988,000 and 311,000, respectively, for the total U.S. The ACE interview workload is over 300,000 housing units because, for budgeting purposes, we want to be conservative and reflect the possibility that we may design for a slightly larger number of housing units. While for Puerto Rico, the comparable numbers are approximately 47,000 and 14,500. The number of housing units designated for person interview includes all occupied and vacant P-sample units. Due to a recent decision in reference 1, supplemental housing units will not be part of interviewing. The estimates of occupied P-sample housing units are provided for documentation. Note that sample size estimates for any of the Person Follow-up and the Housing Unit Follow-up operations are not reflected in any of these housing unit estimates.

TABLE: ACE ESTIMATED SAMPLE SIZES

Sample Size	U.S. and Puerto Rico Total	U.S.			Puerto Rico		
		Total	Medium and Large Blocks	Small Blocks	Total	Medium and Large Blocks	Small Blocks
P-sample Housing Units (Occupied)	285,500	272,600	271,700	900	12,900	12,800	100
P-sample Housing Units (Occupied and Vacant)	325,500	311,000	310,000	1,000	14,500	14,400	100
Listed Housing Units	2,035,000	1,988,000	1,978,000	10,000	47,000	46,700	300
Listed Block Clusters	30,575	30,000	25,000	5,000	575	480	95

The data used to derive these estimates include:

- 1990 Census housing unit and block size distributions
- 1997 annual average state gross vacancy numbers (based on unofficial Housing Vacancy Survey estimates)
- an 11.3 percent vacancy rate (1990 Census) for Puerto Rico with 40 percent of all the housing units being in large blocks
- proportional allocation of sample within state for listing
- the sample of 355 block clusters from American Indian reservations

Note that the number of block clusters that is designated for interview is currently unknown.

Reference

1 Memorandum for J. Thompson from H. Hogan, "Changes Planned for the Accuracy and Coverage Evaluation Interviewing," March 31, 1999, DRAFT.

cc:
DSSD Census 2000 Procedures and Operations Memorandum Series Distribution List
ACE Implementation Team
Statistical Design Team Leaders
Sample Design Team
Ed Kobilarcik (DMD)
Mimi Born

UNITED STATES DEPARTMENT OF COMMERCE
Bureau of the Census
Washington, DC 20233-0001

MASTER FILE

May 3, 1999

DSSD CENSUS 2000 PROCEDURES AND OPERATIONS MEMORANDUM SERIES R--8

MEMORANDUM FOR Robert W. Marx
Chief, Geography Division

From: Howard Hogan *Howard Hogan*
Chief, Decennial Statistical Studies Division

Prepared by: *PPD* Peter P. Davis and Thomas Mule *VTM*
Sample Design Team,
Decennial Statistical Studies Division

Subject: Census 2000 Specifications for Block Cluster Formation–Reissue

This memorandum is being reissued for inclusion in the official *DSSD Census 2000 Procedures and Operations Memorandum Series*. All text in the version of this memorandum distributed on February 16, 1999 remains the same in this version.

I Introduction

As a preliminary stage of the Post-Enumeration Survey (PES) design, this memorandum addresses the formation of block clusters. The PES program for Census 2000 requires that a sample of housing units be selected for intensive reinterviews. The goal of block clustering is to form a group of blocks that average 30 housing units and can be identified by interviewers in the field. PES design consists of dividing the United States into block clusters, groups of geographically contiguous blocks and housing units. Then a sample of block clusters will be drawn. Geography Division (GEO) will perform the clustering. Section II lists the assumptions utilized in the block clustering design. Section III lists the input files/systems that will be used in clustering. Section IV describes the specifications for the formation of block clusters for Census 2000. Section V describes certain identification processes in the block clustering procedure. Section VI lists the desired outputs needed for clustering verification and monitoring future sampling operations. Section VII involves the after sampling file preparation.

These specifications should be used to flowchart the process, to generate further discussion on requirements, to identify and finalize the record layouts of input and output files, and to write computer software to implement the methodology. During and after a testing phase, it is likely that changes to the specifications will be necessary.

Please direct any questions about these requirements to
Peter Davis (email: Peter.P.Davis@ccmail.census.gov) or
Thomas Mule (email: Vincent.T.Mule.Jr@ccmail.census.gov).

II Assumptions

Block cluster formation will take place with these underlying assumptions:

A. The 2000 collection blocks will be clustered.

B. The waves 1, 2, and 3 of Address Listing field operation are completed and will be used. Wave 4 will be done concurrently with Block Canvassing and will not be used. The housing unit (HU) information has been included on the Master Address File (MAF), but the resulting suffixing may not be done.

C. No Block Canvassing results will be used for clustering. Wave 1 of Block Canvassing will not be complete by the start of block clustering.

D. The type of enumeration area (TEA) conversion will not be totally completed.

   1. Most of the Mailout/Mailback to Update/Leave conversion areas will be identified but there will be no current HU counts available. The current HU counts for these areas will not be determined until after Wave 4 of the Address Listing operation is completed. For these identified areas, the 1990 counts will be available. For some areas, this conversion will not happen until after clustering.

   2. The Update/Leave to Mailout/Mailback conversion has been delayed. Clustering will not have any of these conversions.

   3. The Urban Update/Leave conversion is happening now. Additional conversions may occur in subsequent months.

E. Suffixes will not be used to identify individual blocks. For example, blocks 10001A and 10001B will be treated as block 10001 and assigned to the same block cluster.

F. Commercial units and Group Quarters will be properly identified as such and not treated as housing units. Commercial units can be identified on the MAF by the variable RESSTAT. Group Quarters can be identified on the MAF by the variables GQHUFLAG, GQ_NAME, MAFSRC, and GQID.

G. Small block clusters have 0, 1, or 2 HUs. Medium block clusters have between 3 and 79 HUs. Large block clusters have over 79 HUs.

H. There is insufficient information to identify Crews on Vessels areas to exclude them from the block clustering process.

I. The most recent Delivery Sequence File (DSF) update was September, 1998. The next DSF update is scheduled for April, 1999.

J. "Inside the blue line" refers to Mailout/Mailback TEA areas, that is, areas with predominantly city-style addresses. "Outside the blue line" refers to the remaining TEAs, areas with predominantly non-city-style addresses.

K. The Decennial Master Address File (DMAF) extraction criteria used to define MAF housing unit counts is the ideal in the block clustering process. However, the DMAF extraction criteria have not been determined at the time of clustering. MAF housing unit criteria is specified based on a possible set of rules being considered.

L. Input files, systems, and results are benchmarked at the time of clustering. Future changes to input files and systems are expected, so discrepancies could occur if comparing clustering results.

M. A complex algorithm, which is assumed correct, is implemented to identify military blocks. This algorithm gathers information from several sources within TIGER to properly identify military blocks.

N. During the actual process of forming block clusters, blocks are considered neighbors if they share a line segment boundary. Two blocks joining at a point are not neighbors.

O. If two (2) or more kinds of boundaries separate adjacent blocks, then the following order of boundaries will have priority:

1. Boundaries listed in Section IV, Paragraph B, Step 1. These boundaries will never be crossed.
2. Invisible boundaries as defined in Appendix C. These block clusters will be collapsed together.
3. Areal water ("two-line water"), limited access highway or ridge line. These boundaries will never be crossed.
4. Streams ("one-line water") or rail lines.
5. Any remaining boundary.

III Input Files/Systems

The following files/systems will be used in the Census 2000 clustering process.

A. Most recent MAF: This file should include the updates from waves 1, 2 and 3 of Address Listing operations that were just conducted. The last DSF update was September, 1998. No Census 2000 LUCA actions (adds, deletes or corrections) will be reflected since the verification will not be done by the time of clustering. No Block Canvassing results will be used.

B. TIGER system: The TIGER (Topologically Integrated Geographic Encoding and Referencing) system provides proximity, perimeter, and area information for the Census 2000 collection blocks. TIGER itself is a cartographic database with physical features (such as roads, railroads, and rivers) and address ranges. TIGER also contains 1990 housing unit estimates.

IV Block Cluster Specifications

A. Housing Unit Counts

1. Defining/Tallying MAF Housing Unit Counts

MAF housing unit counts are required when creating block clusters. The official definition of a MAF housing unit has not yet been finalized. The MAF contains many different types of addresses and address sources. The variable UNITSTAT from the MAF represents the characteristic of an address used in identifying housing units for the purposes of block clustering. Commercial Units and Group Quarters will be excluded using the MAF variables RESSTAT, GQHUFLAG, GQ_NAME, MAFSRC, and GQID.

The UNITSTAT codes that will be recognized as valid housing units for block clustering are as follows:

1 = Valid Living Quarters
6 = Under Construction
8 = Vacant Trailer Pad
10 = Boarded Up
11 = Unable to locate
12 = Seasonal
13 = Vacant
15 = Map Spotted Unit with insufficient information*

* A unit with insufficient information includes those addresses outside the blue line that are missing map spot numbers or that have neither a mailing address nor a location description. Include as a valid housing unit those units that have either 1) a mailing address, or 2) a map spot number and a location description.

A complete list of all MAF housing unit codes can be found in Appendix H.

Along with UNITSTAT, field operations will also be used to identify MAF housing units for block clustering as follows:

- For Address Listing areas, an address with one of the UNITSTAT codes listed above will be tallied as a housing unit for block clustering if the address was field verified during Address Listing.

- For areas outside Address Listing, including the three Dress Rehearsal sites, an address with one of the UNITSTAT codes listed above will be tallied as a housing unit for block clustering, including:

    1. addresses whose sole source is the 1990 Address Control File (ACF),
    2. addresses that appear in both the ACF and DSF, and also appear in the most recent DSF,
    3. addresses that appear in both the ACF and DSF, but do not appear in the most recent DSF,
    4. addresses whose sole source is the DSF, and also appear in the most recent DSF,
    5. addresses whose sole source is the DSF, but which do not appear in the most recent DSF.

In addition, for the three Dress Rehearsal sites, include addresses and address actions resulting from the following field operations:

1. Be Counted/Telephone Questionnaire Assistance (TQA)
2. Nonresponse Follow-Up (NRFU)
3. Local Update Census Address (LUCA) Field Verification
4. Targeted Canvass (inside the blue line)
5. Targeted Multi-unit Check (inside the blue line)
6. Update/Leave (outside the blue line)
7. U.S. Postal Service Casing Check

This definition of a MAF HU is intended to be as similar as possible to the final DMAF extraction criteria. However, since the final definition is not known, differences between these two definitions may occur.

2. Determining Housing Unit Counts for Clustering

For block clustering, the most up-to-date counts available will be used. In the past year, Census Bureau personnel performed waves 1, 2 and 3 of the Address Listing operation in most areas outside the blue line; that is, in areas with predominantly non-city-style addresses. The resulting HU counts are considered the most up-to-date. List/Enumerate areas and blocks converted from Mailout/Mailback to Update/Leave were not included in waves 1, 2 or 3 of the Address Listing operation. The only available count for these areas is the 1990 housing unit count. For the remaining blocks, the benefit of Block Canvassing will not be available. No Census 2000 LUCA actions (adds, deletes or corrections) will be used in the block clustering.

Obtain the HU counts to be used for clustering as follows:

a. For blocks included in waves 1, 2 or 3 of the Address Listing operation which include Update/Leave blocks, Rural Update/Enumerate and Update/Leave to Mailout/Mailback conversion blocks: use the housing unit counts from the most recent MAF. These counts were determined in the past year by Census Bureau personnel and are considered the most up-to-date housing unit counts.

b. For List/Enumerate blocks and Mailout/Mailback to Update/Leave conversion blocks: use the 1990 HU counts. In these areas, no other counts are available.

c. For all remaining blocks which include Mailout/Mailback blocks, Urban Update/Leave blocks, and Military blocks: use the correspondence between 1990 tabulation blocks and 2000 collection blocks to determine which HU count to use as follows:

   i. One-to-one or a many-to-one correspondence between 1990 tabulation block and 2000 collection block: use the higher of the most recent MAF or 1990 count.

   ii. One-to-many correspondence between 1990 tabulation block and 2000 collection block: use the most recent MAF count. For the one-to-many correspondence, the 1990 counts in TIGER are estimated to the 2000 blocks based on land area. The most recent MAF gives a more up-to-date number in this instance.

These rules mean that it is possible that an individual block cluster may contain a combination of 1990 and MAF counts. For a summary of HU allocation based on Type of Enumeration Area, see Appendix I.

3. Tracking Housing Unit Counts

Keeping track of HU counts may provide information that could explain any discrepancies during the sampling process.

a. Place both the most recent MAF HU counts and the 1990 HU counts on the block cluster output file. The layout for the file is in Appendix A.

   For both fields, assign the following:

   If the source has a HU count, assign the HU count (0-99999)
   Else if the source does not have a count, assign a blank ('   ').

   Note: An example of when a source does not have a HU count is List/Enumerate area where there is no MAF count.

b. Place an indicator variable for each block on the output file. It will indicate if the housing unit count used for clustering comes from the most recent MAF or the 1990 counts.

For this field, assign the following:

If the count is from the most recent MAF, assign the Housing Unit Count Indicator = '1'
Else if the count is from the 1990 count, assign the Indicator = '2'.

B. Formation Rules

This formation is hierarchical. That is, step 1 takes precedence over step 2 and so on.

1. Block clustering will adhere to several geographical constraints. Block clusters will not cross the following boundaries:

   a. County and sub-county partitions in the TIGER system: counties will be a boundary. For large counties, The TIGER system maintains county parts on separate files. These county parts are referred to as sub-county partitions. In these counties, a block cluster will not cross the sub-county partition. A list of these counties is given in Appendix B.

   b. Census Tract: PES block clusters will respect the 6 digit (including suffixes) tract boundaries for 2000. The tract definitions GEO will use to cluster are the interim tract boundaries. Interim tracts are the 1990 tracts adjusted for the 2000 collection block boundaries and are also referred to as pseudo-tracts. Census 2000 tracts will not be defined until some time in 2000.

c. Groups of Type of Enumeration Areas (TEA): The following TEAs may be clustered together.

TEA Group	Description	TEA
A	Mailout/Mailback	1
	Urban Update/Leave	7
	Update/Leave to Mailout/Mailback conversions	8
B	Update/Leave	2
	Rural Update/Enumerate	5
C	List/Enumerate	3
D	Military (outside the blueline)	6
E	Mailout/Mailback to Update/Leave conversions	9

d. Military Areas: military blocks should be clustered with military blocks. For military areas classified as Military Type of Enumeration Area, this is accomplished by the TEA restriction, above. A complex algorithm, assumed to be correct, gathers a set of TIGER information used to identify military blocks inside the blue line.

e. American Indian Country (AIC) Land: Block clusters on AIC land will respect the American Indian Reservation/Alaska Native Village Statistical Area (ANVSA) values from TIGER. The values are as follows:

AIR/ANVSA Values	Element Description
Blank	No American Indian Country specified
0001-4989	American Indian reservation/trust land
5001-5989	Tribal jurisdiction statistical area
6001-5989	Alaska Native Village Statistical Area
9001-9589	Tribal designated statistical area

i. American Indian Reservation (AIR) blocks are allowed to be clustered with other AIR blocks on the same reservation.

ii. Blocks on AIC but outside AIR are allowed to be clustered with other AIC but outside AIR blocks.

iii. The blocks must have the same ANVSA value to be combined.

iv. If a block is partially on an American Indian Reservation, classify the block as AIR. If the block is partially on an American Indian Country and not in an AIR then classify as AIC.

v. Blocks that contain two AIRs are assigned only one AIR value. The rule for designating these blocks to only one AIR value is accomplished in a complex algorithm designed by the TIGER Systems Branch and is assumed correct.

2. Exclude the following blocks from the clustering procedure. They will not be clustered with any adjacent blocks, and they will not be included on the verification and the block cluster output files sent to Decennial Systems and Contracts Management Office (DSCMO).

   a. Blocks consisting entirely of bodies of water (water blocks)
   b. Blocks in Remote Alaska [TEA=4]

3. All blocks within each county separated by invisible boundaries will be collapsed except when the invisible boundary is a block extension identified by Census feature codes F20 (Feature Extension–extensions not otherwise classified), F21 (Automated extension), or F22 (Irregular block extension). All Feature Class F codes besides F20, F21 and F22 will be collapsed. Intermittent streams (Feature Code H12) and Intermittent canals, ditches, or aqueducts (Feature Code H22) forming block boundaries will also be considered to be invisible boundaries. See Appendix C for the invisible boundary codes.

4. All blocks with 80 or more housing units will be block clusters by themselves. No other blocks will be clustered with this block unless the block completely surrounds another block with zero HUs. (see rule #9.)

5. Any block larger than 15 square miles will be a block cluster by itself unless the block cluster completely surrounds another block. (see rule #9.)

6. Block clusters that are not contiguous with any other block clusters within the boundaries specified in step 1 will be block clusters by themselves.

7. Follow these guidelines in hierarchical order when combining neighboring blocks:

   a. Never cross areal water ("two-line water"), limited access highways, and ridge lines.
   b. Do not create a block cluster with 80 or more housing units.
   c. Do not create a block cluster with more than 15 square miles. (Exception: see Rule #9.)

d. Never cross streams ("one-line water") or rail lines except when there are no other adjacent clusters available for clustering. If a cluster must cross either a stream or a rail line, it should cross the stream before the rail line.

8. Sometimes, a block can completely surround another block. The block which is enclosed within the surrounding block can be identified because it has only one neighbor. If a block has only one neighbor, combine as follows:

   a. If the surrounding block has fewer than 80 housing units (even blocks with 0 housing units), then collapse the surrounding block with the enclosed block as long as the resulting block cluster does not exceed 80 HUs.

   b. If the surrounding block has 80 or more housing units, then:

      i. collapse the surrounding block cluster with the enclosed block only if the enclosed block has a total of 0 housing units.

      ii. otherwise, do not combine.

9. All blocks with more than 15 square miles:

   a. are NOT eligible to initiate the clustering algorithm.
   b. are NOT considered eligible neighbors for clustering.
   c. are NOT eligible to be combined when they are surrounded by another block.
   d. are eligible to be combined with a block when it surrounds another block (see rule #9).
   e. are NOT eligible to absorb block clusters with 0 housing units along the perimeter.

10. All blocks containing 3 to 25 housing units:

   a. will initiate the clustering algorithm.
   b. are eligible to be collapsed with neighboring blocks such that the resulting cluster does not exceed 80 HUs.
   c. are eligible to be combined when they are surrounded by another block (see rule #9).
   d. are eligible to be combined with a block when it surrounds another block (see rule #9).
   e. are eligible to absorb block clusters with 0 housing units along the perimeter.

11. All blocks containing 26 to 79 housing units:

    a. do NOT initiate clustering.
    b. are NOT considered eligible neighbors for clustering.
    c. are eligible to be combined when they are surrounded by another block (see rule #9).
    d. are eligible to be combined with a block when it surrounds another block (see rule #9).
    e. are allowed to absorb blocks with 0 housing units along the perimeter.

12. All blocks with 1 or 2 housing units:

    a. do NOT initiate the clustering algorithm.
    b. are eligible for collapsing with neighboring block clusters only if the adjacent block cluster contains at least 3 housing units.
    c. are eligible to be combined when they are surrounded by another block (see rule #9).
    d. are eligible to be combined with a block when it surrounds another block (see rule #9).
    e. are allowed to absorb blocks with 0 housing units along the perimeter.

13. All blocks with 0 housing units:

    a. are NOT eligible to initiate the clustering algorithm.
    b. are NOT eligible to be a neighboring block cluster.
    c. are eligible to be combined when they are surrounded by another block (see rule #9).
    d. are eligible to be combined with a block when it surrounds another block (see rule #9).
    e. are NOT eligible to absorb block clusters with 0 housing units along the perimeter.

14. All blocks with 80 or more HUs:

    a. are NOT eligible to initiate the clustering algorithm.
    b. are NOT eligible to be a neighboring block cluster.
    c. are NOT eligible to be combined when they are surrounded by another block (see rule #9).
    d. are eligible to be combined with a block when it surrounds another block (see rule #9).
    e. are NOT eligible to absorb block clusters with 0 housing units along the perimeter.

Note: See Appendix E for table summaries of Formation Rules 8 through 14.

C. Algorithm

The goal of the algorithm is to cluster blocks and generate clusters that have on average 30 HUs per block cluster for block clusters with 3 to 79 housing units. First, blocks are prepared and screened to identify which phases of processing the block will undergo. There are four basic phases of processing: 1) calculating target size, 2) clustering, 3) checking for enclosed blocks, and 4) a zero block perimeter search. The algorithm proceeds as follows.

1. Calculate the Clustering Target for the county

The overall goal of clustering is to have an average of 30 HUs per cluster for the medium clusters. Since many blocks have more than 30 HUs, if clusters of size 30 are formed, the overall average will be greater than 30. Therefore, a clustering target for blocks containing 3 to 25 HUs which balances the number of blocks greater than or equal to 26[1] is determined. Do this as follows:

a. For each county, classify blocks that have between 3 and 79 HUs into two groups: 1) between 3 and 25 HUs inclusive and 2) between 26 and 79 HUs inclusive. For each group, count the number of blocks and the number of HUs. Blocks that are between 26 and 79 HUs will be considered block clusters by themselves. This can be used to determine at what average does the algorithm need to cluster the 3 to 25 HU blocks so that the overall medium block cluster average is as close to 30 as possible.

Calculate the Clustering Target:

$$\text{Cluster Target} = \frac{(\#\text{ of HU}_{3\text{-}25})}{\left[\left(\dfrac{\#\text{ of HU}_{3\text{-}25} + \#\text{ of HU}_{26\text{-}79}}{30}\right) - \#\text{ of blocks}_{26\text{-}79}\right]}$$

If the number of blocks containing 3 to 25 HUs is zero, then proceed to Step 2, Early Stages of Block Clustering. If the Cluster Target is less than 10, set the Cluster Target to 10. If the sum of the number of HUs between 3 to 25 plus the number of HUs between 26 to 79, divided by 30, equals the number of blocks containing 26 to 79 HUs, then set the Cluster Target to 26.

---

[1] 26 is used instead of 30 because 26 is within 15% of 30. 15% is considered sufficiently close to the target to stop clustering.

Note: This average is calculated before the invisible boundaries are collapsed.

b. For example, a county has 5479 blocks that have between 3 and 25 HUs per block. In these 5479 blocks, there are 77,696 HUs. There are 4087 blocks that have between 26 and 79 HUs per block. In these 4087 blocks, there are 167,103 HUs.

The cluster target would be calculated as follows:

$$\text{Cluster Target} = \frac{77696}{\left[\left(\frac{77696 + 167103}{30}\right) - 4087\right]} = 19.1$$

The 3 to 25 HU blocks would be clustered with a desired average of 19.1 HUs per block cluster. This will produce a medium block cluster average that is close to the goal of 30 HUs per block cluster.

c. This average is calculated separately for each county processed. Since counties can have diverse block HU density distributions, the medium block cluster average goal of 30 HUs is better achieved by having each county's medium block clusters average around 30.

d. Blocks with 1 or 2 HUs are eligible to be combined if they are the closest neighbor. These HUs were not used in computing the above average because 1) the number of 1 or 2 HU blocks to be combined is unknown and 2) the total number of HUs in blocks with 1 or 2 HUs is very small as compared to the total number of HUs in the medium blocks.

2. Early stages of block clustering: Prepare and Screen the Blocks

a. The algorithm checks the geographic constraints first. These boundaries are never crossed. The next step is to collapse the blocks that are separated by invisible boundaries.

b. The 80+ HU blocks and the 15 sq. mile blocks are separated and considered clusters themselves, not eligible to initiate clustering. Block clusters containing either 80+ HUs are eligible to be collapsed if they surround a block containing zero HUs. Blocks spanning more than 15 sq. miles are eligible to be collapsed only if they completely surround a block and meet the housing unit requirements (Formation Rule #9). No other block clusters will be clustered with either an 80+ HU block or a 15 sq mile block.

c. Blocks between 26 and 79 HUs are considered to be clusters by themselves, not eligible to initiate clustering. These block clusters are eligible to be collapsed if they 1) completely surround another block or another block completely surrounds it and 2) satisfy the enclosure rules (Formation Rule #9). These blocks are eligible to perform the zero neighbor perimeter search.

d. The small blocks (0, 1, or 2 HUs) are set aside for the time being as individual clusters. They will be eligible to be clustered later on. Any that are not collapsed remain as small block clusters.

3. Form block clusters:

a. This stage begins with the blocks containing 3 - 25 HUs in the TIGER order. All blocks with the number of housing units greater than 85% of the cluster target, calculated in part 1, are considered to be sufficiently close to the preferred average block cluster HU size and although they do not initiate clustering, they are eligible neighbors for collapsing.

b. Given a block, call it block A, that has fewer HUs than 85% of the cluster target, identify an eligible list of neighbors. Collapse block A with its closest neighbor. The closest neighbor is defined by the block with the closest centroid that has at least 1 but no more than 25 HUs and a shared line segment boundary.

c. Let block A's closest neighbor be block B. A new block cluster, call it AB, is formed by the combination of Block A and its neighboring block, block B. If the total number of housing is greater than 85% of the cluster target then proceed to Step 4: Enclosed Blocks. If the total number of housing units is less than 85% of the cluster target then find the closest neighbor, block C, and form ABC. Continue to find the closest neighbor and collapse it into the block cluster until the total number of housing units is greater than 85% of the cluster target.

d. Once a block cluster reaches 85% of the cluster target, it is written off to the used file and removed from the neighbor list. This block cluster will not be an eligible neighbor for any ensuing blocks in the TIGER order. The newly formed block cluster will then proceed to the enclosed block and perimeter search steps.

4. Collapse Enclosed Blocks:

   a. Definition: Enclosed blocks are blocks with only one neighbor.

   b. Collapse the enclosed block with its surrounding block if the enclosed block has < 80 HUs and the resulting cluster is not > 80 HUs. If the enclosed block has zero units, it can be combined with an 80+ HU block.

5. Zero Neighbor Perimeter Search:

   Once a block cluster has been collapsed with its closest neighbors to contain more than 85% of the cluster target and checked for enclosed blocks, the final step involves searching the perimeter of the block cluster. If there is a neighboring block on the perimeter that has 0 housing units, then the 0 housing unit block is to be collapsed into the block cluster.

6. Proceed to the next block that has fewer HUs than 85% of the cluster target, and restart the clustering process at step 3b, above.

V Identification

A. Block Cluster Number

Within county, each block cluster will be uniquely identified based on a numbering process which uses the cluster's latitude and longitude. The latitude and longitude are merged, sorted, and then transformed into a 5-digit GEO cluster number. This cluster number will produce a geography sort.

B. Collapsing Across Invisible Boundaries

Block clusters that were formed by collapsing invisible boundaries need to be identified (see Appendix C.) A block cluster may have 1) more than 1 block and 2) more than 80 HUs or 15 square miles. One reason for this occurrence is if invisible boundaries are collapsed. The indicator variable will allow us to verify this.

Assign the Invisible Boundary Indicator field to the output file:

1. If an invisible boundary is crossed, assign a value of '1'.
2. If no invisible boundaries are crossed, assign a value of '0'.

C. American Indian Country Block Clusters

Block clusters that contain American Indian Country need to be identified. These are lands that are American Indian Reservation/trust land, tribal jurisdiction statistical area, tribal designated statistical area, and Alaska native village statistical area. American Indians will have their own sampling stratum in 2000 and hence need their own identification for the clustering process.

Use the variable American Indian Reservation/Alaska Native Village Statistical Area (ANVSA) to identify the types of American Indian Country. This variable is defined in TIGER Documentation: Chapter III, Section B, TIGER System County Partition Data Element Definitions. Assign the American Indian Country field to the output file as follows:

AIC Indicator	ANVSA Values	Element Description
0		No American Indian Country specified
1	0001-4989	American Indian reservation/trust land
2	5001-5989	Tribal jurisdiction statistical area
	6001-5989	Alaska Native Village Statistical Area
	9001-9589	Tribal designated statistical area

D. Type of Enumeration Area

Put two Type of Enumeration Area codes on the cluster output file. The first is the Initial Block TEA [values: 1, 2, 3, 4, 5, 6, 7, 8 or 9]. This is the block TEA value at the time clustering occurred. The second is the TEA Group variable [values: A, B, C, D, or E (See Formation Rule 1c)]. This is the TEA group value that is assigned at the time of clustering.

E. Military Area Indicator

Blocks on Military Areas need to be identified. Military areas are a boundary in step 1 of the Formation Rules. Put an indicator on the output file to denote if the block is a military area. This provides a check that the rule was implemented correctly. Outside the blue line, military areas are designated by TEA value of 6. For military blocks inside the blue line, Geography division will use TIGER to identify them.

Assign the Military Area Indicator to the file as follows:

1. If no military area in block, then assign value of '0'.
2. If block contains military area then assign value of '1'.

VI Output

A. Equivalency Files

A correspondence between 1990 Tabulation and 2000 collection blocks is required so that DSCMO can determine the approximate demographic composition of each block cluster. Make available the standard Block Equivalency File in the standard format relating the 2000 collection blocks to the 1990 tabulation blocks. The format for the equivalency file is located in Appendix J.

B. Verification Maps

GEO will deliver block cluster maps for review by the Decennial Statistical Studies Division (DSSD) for use in verifying the block clustering for Census 2000. There will be one multi-page map for each county requested. (See section D., Testing and Production, below.) At a minimum, these maps should include block boundaries, block numbers, block cluster boundaries, and block cluster numbers. Also included on the maps will be the following color designations for these respective Type of Enumeration Areas: TEA 1, TEA 7, and TEA 8 should be similar shades of blue, TEA 2 and TEA 5 should be similar shades of red, TEA 3 should be yellow, TEA 6 should be green, and TEA 9 should be purple.

C. Block Files

The block clustering operation will be a flowing process involving GEO, DSSD, and DSCMO. The verification process will be based upon reviewing maps, verification files and summary statistics. Once the clustering file results have been reviewed and approved for a state then the completed files will be provided from GEO to DSCMO. DSSD will give official approval for GEO to make files available to DSCMO.

1. Verification Files

The files delivered by GEO to DSSD for Census 2000 will be similar to that of the cluster files for the 1998 Dress Rehearsal. Changes have been made to add the indicator fields created and the housing unit count information.

A selected number of counties for each state will be specified for review by DSSD. Each file will contain one record for each block for the selected counties in the state. (See section D., Testing and Production.)

2. File to DSCMO

After a state has been verified, GEO will deliver the complete file (all blocks in that state) to DSCMO staff. The file should contain all of the fields listed in the block cluster verification file layout. (See Appendix A.)

Place the file in the GEBA01::GEO_PUBLIC:[PES.BC] subdirectory. DSCMO will be able to obtain the file from there.

D. Testing and Production

For testing purposes, several counties from the initial stages of block clustering will be reviewed. This testing procedure will examine these counties to check the rules, the algorithm, and the overall goal of block clustering. This procedure will occur prior to production.

DSSD will use the maps and GusX to visually inspect the block clusters. Using the verification output file, a SAS program will be written to identify any discrepancies in the rules, algorithm, and/or goal of block clustering. The TIGER Systems Branch will generate a Cluster Summary File that contains the number of HUs, the number of blocks, the number of zero HU blocks, the Cluster Target, the number of HUs between 3 and 25, the number HUs between 26 and 79, and the number of blocks between 26 and 79 for each cluster.

For testing, the Dress Rehearsal sites will be run and made available for review. Also, DSSD will examine one county from each of the Wave 1 states. The counties that will be tested are as follows: Washington, DC, Hennepin County, MN, Denali Borough, AK, Glacier, MT, Mountrail, ND, Shannon, SD, Florence, WI, and Washington, ID. For two counties, GEO will produce maps. For the remaining counties, DSSD staff can visit GEO and use GusX to review these counties instead of producing maps.

During production, DSSD will examine two counties from each of the Wave 1 states. The verification and Cluster Summary File for each county and the state summary file will be sent to DSSD after the state is processed. The Cluster Summary File only needs to be produced for the production check counties. One county from each state is examined during testing. If DSSD compares the testing to the production files in these counties, no differences should exist. Hence, no maps for these counties are necessary. Maps will only be needed for the other counties.

DSSD will examine one county from each of the Wave 2 states. Maps, verification file, the Cluster Summary File and the state summary file will be sent to DSSD.

For Wave 3 and 4 states, one county in California, Illinois and Texas will be completely reviewed. Maps, the county verification file, the Cluster Summary File and the state summary file will be sent to DSSD. For the remaining states and Puerto Rico, one county has been chosen to be computer checked. The county verification file, Cluster Summary File and state summary file will be sent to DSSD. No maps need to be made for these counties.

Appendix F contains the order in which the states will be processed during testing and production. Appendix G contains the counties to be verified during production.

E. Summary Counts

GEO will produce summary block cluster counts for DSSD to use in monitoring future sampling operations. Also, it will allow DSSD to examine summary results of the clustering in every county/partition in the state during verification. Having these at the county/partition level will allow DSSD to combine them as needed. The Cluster Target calculated for each county/partition will also be allocated to this file. This will document the cluster target used for each county/partition.

This file will be an ASCII file for each state with one record per partition and will be sent with the verification materials. Unlike the verification files, all counties/partitions will be listed on this file. The layout of the file is in Appendix D.

VII After Sampling File Preparation

GEO will send the block cluster files to DSCMO, which will then select a sample of clusters. After sampling, DSCMO will send a file of the sampled clusters to GEO so that three fields can be added to the file. GEO will assign 1) revised block TEA, 2) revised city-style address indicator and 3) the Local Census Office (LCO) code. This information will be used in the creation of the Collection Geographic Reference File (GRF). Return the updated sample file to DSCMO.

A. **Revised Block TEA**

The block TEA values may change after the block clustering process is completed. After the sample is selected, the TEA codes for the sampled block clusters need to be updated.

Assign the Revised Block TEA (RBTEA) value for each block.

B. **Cluster TEA Code**

PES operations handle block clusters that have city-style addresses differently than block clusters with non-city-style addresses. Because of this, a Cluster TEA code is needed to identify city-style address clusters and non-city-style address clusters. Clusters that have at least 1 block that is Update/Leave (RBTEA=2), List/Enumerate (RBTEA=3), Rural Update/Enumerate (RBTEA=5), or Block Canvassing moved to Address Listing (RBTEA=9) are considered to be non-city-style address clusters. Clusters that contain solely blocks that are Mailout/Mailback (RBTEA=1, 6 or 8) or Urban Update/Leave (RBTEA=7) are considered to be city-style address clusters. This is a cluster-level variable. Each block in the cluster will receive the same cluster value. Assign the city-style address indicator to each block as follows.

If a cluster has at least one block with an RBTEA = 1, 7 or 8 then assign the Cluster TEA Code = '2' (non-city-style address) to all blocks in the cluster; otherwise, assign the Cluster TEA Code = '1' to all blocks in the cluster.

C. **Local Census Office (LCO) Code**

After the PES sample is selected, the Census LCO field will need to be updated. For each block on the file, assign the LCO code to the file.

cc: DSSD Census 2000 Memorandum Series Distribution List
PES Implementation Team
Statistical Design Team Leaders
Sample Design Team
C. Hantman (GEO)
R. Ruiz (GEO)
S. Holt (GEO)
K. Todd (GEO)

Appendix A
Verification and DSCMO Block Cluster File Layout

The following is the layout of the block clustering file that will be made available to DSCMO. The verification files will include all of the records for the counties specified by DSSD. There will be a record for each block. When a state is verified, all of the county files will be combined into one file and made available to DSCMO.

Variable	Location
State	1:2
County	3:5
Interim Tract (a.k.a. pseudo-tract)	6:11
Block Number	12:16
Blank	17:17
Cluster Number	18:22
Blank	23:23
Cluster Size code 1 = Clusters with 0 HUs 2 = Clusters with 1 HUs 3 = Clusters with 2 HUs 4 = Clusters with between 3 and 5 HUs 5 = Clusters with between 6 and 9 HUs 6 = Clusters with between 10 and 19 HUs 7 = Clusters with between 20 and 29 HUs 8 = Clusters with between 30 and 79 HUs 9 = Clusters with 80 or more HUs	24:24
Blank	25:25
Block Area (Sq. Miles)	26:31
Blank	32:32
Block Perimeter (Miles)	33:37
Blank	38:38
Block Cluster Area (Sq. Miles)	39:44
Blank	45:45
Block Cluster Perimeter (Miles)	46:50
Number of HUs in cluster	51:55
Number of HUs in block	56:60

Variable	Location
Block TEA	61:61
1 = Mailout/Mailback	
2 = Update/Leave	
3 = List/Enumerate	
5 = Rural Update/Enumerate	
6 = Military	
7 = Urban Update/Leave	
8 = Update/Leave to Mailout/Mailback conversions	
9 = Mailout/Mailback to Update/Leave conversions	
TEA Group for Block Cluster	62:62
A= Mailout/Mailback or Update/Leave to Mailout/Mailback conversions	
B= Update/Leave or Rural Update/Enumerate	
C=List/Enumerate	
D=Military	
E=Mailout/Mailback to Update/Leave conversions	
2000 MAF HUs count	63:67
' ' Blank if no HU count available	
1990 ACF HUs count	68:72
' ' Blank if no HUs count available	
Housing Unit Count Indicator	73:73
1 = from 2000 MAF	
2 = from 1990 ACF	
Invisible Boundary Collapse Indicator	74:74
0 = No	
1 = Yes (Collapsing across Invisible Boundary in BC)	
American Indian Country Indicator	75:75
0 = No American Indian Country	
1 = American Indian Reservation/trust land	
2 = Tribal jurisdiction statistical area/ Alaska Native Village statistical area/ tribal designated statistical area	
Military Indicator	76:76
0 = No Military Area	
1 = Block contains Military Area	

Appendix B
Partitioned Counties in the TIGER system

The following counties are partitioned into smaller sub-county geographic areas in the TIGER system. The block clustering in these counties will have to observe these sub-county partition boundaries.

04013 Maricopa, AZ
04019 Pima, AZ
06029 Kern, CA
06037 Los Angeles, CA
06059 Orange, CA
06065 Riverside, CA
06071 San Bernardino, CA
06073 San Diego, CA
06085 Santa Clara, CA
12025 Dade, FL
12099 Palm Beach, FL
15003 Honolulu, HI
17031 Cook, IL
17043 DuPage, IL
25017 Middlesex, MA
26163 Wayne, MI
36059 Nassau, NY
36103 Suffolk, NY
39049 Franklin, OH
42003 Allegheny, PA
48029 Bexar, TX
48113 Dallas, TX
48157 Fort Bend, TX
48201 Harris, TX
48439 Tarrant, TX
53033 King, WA

Appendix C
Block Cluster Invisible Boundary Codes

The following is the list of boundary codes that are to be treated as invisible boundaries during the clustering process.

- F00 - Nonvisible boundary, classification unknown or not elsewhere classified
- F10 - Nonvisible governmental unit boundary
- F11 - Offset corporate boundary
- F12 - Corporate corridor
- F13 - Nonvisible interpolated boundary
- F14 - Superseded political boundary
- F15 - Corrected governmental unit boundary
- F16 - EAC nonvisible boundary
- F17 - State legislative non-visible boundary
- F18 - Congressional District non-visible boundary
- F23 - Closure extension
- F24 - Nonvisible separation line
- F25 - Nonvisible corporate corridor centerline
- F30 - Point-to-point line
- F40 - Property line
- F50 - ZIP Code boundary
- F60 - Map edge
- F70 - Statistical boundary
- F71 - 1980 statistical boundary
- F72 - 1990 block boundary
- F73 - Urbanized area land use boundary
- F74 - 1990 Statistical Boundary
- F80 - Nonvisible other tabulation boundary, major category used when the minor category could not be determined
- F81 - School district boundary
- F82 - Special census tabulation boundary
- F83 - Census 2000 Collection Block Boundary
- F84 - Census 2000 Statistical Area Boundary
- F85 - Census 2000 Tabulation Block Boundary
- F86 - Local Administrative Line

- H12 - Intermittent streams or wash
- H22 - Intermittent canal, ditch, or aqueduct

Appendix D
Summary File Layout

The following is the layout for the partition-level summary statistics requested for each state. There will be one record for each partition. In most states, the partition is equivalent to the county. However, in some states, a county has been subdivided into partitions. (See Appendix B. for counties that are partitioned.) Hence, a partition-level summary is necessary.

Variable	Location
Partition	1:1
State (FIPS code)	3:4
County (FIPS code)	6:8
Number of Medium BCs (Non-Indian)	10:15
Number of Large BCs (Non-Indian)	17:22
Number of Small BCs (Non-Indian)	24:29
Number of Medium BCs with AIR but no other Indian Country land	31:36
Number of Medium BCs with some Indian Country but no AIR land	38:43
Number of Large BCs with AIR but no other Indian Country land	45:50
Number of Large BCs with some Indian Country land but no AIR land	52:57
Number of Small BCs with AIR but no other Indian Country land	59:64
Number of Small BCs with some Indian Country but no AIR land	66:71
Number of 'Water' and Remote Alaska Blocks removed	73:78
Number of HUs on Medium BCs (Contain No American Indian Land)	80:85
Number of HUs on Large BCs (Contain No American Indian Land)	87:92
Number of HUs on Small BCs (Contain No American Indian Land)	94:99
Number of HUs on Medium BCs on American Indian Reservations only	101:106
Number of HUs on Large BCs on American Indian Reservations only	108:113
Number of HUs on Small BCs on American Indian Reservation only	115:120
Number of HUs on Medium BCs on American Indian Country land but not an American Indian Reservation.	122:127
Number of HUs on Large BCs on American Indian Country land but not an American Indian Reservation.	129:134
Number of HUs on Small BCs on American Indian Country land but not an American Indian Reservation	136:141
Number of HUs on "water" or Remote Alaska blocks	143:148
Average Number of HUs per BC for BCs with 3 or more HUs	150:155
Average Number of HUs per BC for BCs with 3 to 79 HUs	157:162
Cluster Target	164:169

Appendix E
Summary of Formation Rules 8 - 14

Block Area (Sq. Miles)	Initiate Clustering?	Eligible to collapse with neighbors?	Eligible to be absorbed if completely surrounded ?	Eligible to check if completely surrounds a block?	Absorb zero blocks along perimeter?
15 +	No	No	No	Yes[1]	No
Less than 15	Follow HU Rules Below				

[1] Not to exceed 80 HUs.

Block Size (HUs)	Initiate Clustering?	Eligible to collapse with neighbors?	Eligible to be absorbed if completely surrounded ?	Eligible to check if completely surrounds a block?	Absorb zero blocks along perimeter?
0	No	No	Yes[1]	Yes[1]	No
1	No	Yes[1,3]	Yes[1]	Yes[1]	Yes
2	No	Yes[1,3]	Yes[1]	Yes[1]	Yes
3 - 25	Yes	Yes[1,3]	Yes[1]	Yes[1]	Yes
26 - 79	No	No	Yes[1]	Yes[1]	Yes
80+	No	No	No	Yes[2]	No

[1] Not to exceed 80 HUs.
[2] Collapse only if there are no housing units in the enclosed block.
[3] Not to exceed 15 Square Miles.

Appendix F
The Production Process of States by Wave

For production, states will be processed in the order listed below. It is anticipated review will follow a similar order. Wave 1 begins with Alaska as the first state to be verified. Once work on Alaska is completed, Idaho is the next state. The production process continues through Wave 4.

Wave 1	Wave 2	Wave 3	Wave 4
Alaska	Arkansas	Delaware	Puerto Rico
Idaho	Connecticut	Maryland	
Minnesota	Hawaii	Ohio	
Montana	Kentucky	South Carolina	
North Dakota	Louisiana	Alabama	
South Dakota	Massachusetts	Florida	
Washington, DC	Mississippi	Georgia	
Wisconsin	Rhode Island	Illinois	
Wyoming	Tennessee	Indiana	
	West Virginia	Iowa	
	Arizona	Kansas	
	Colorado	Michigan	
	Maine	Missouri	
	Nebraska	New Jersey	
	Nevada	North Carolina	
	New Hampshire	Oklahoma	
	New Mexico	Texas	
	New York	Virginia	
	Oregon	California	
	Pennsylvania		
	Utah		
	Vermont		
	Washington		

Appendix G
Production Counties for Review

Wave 1

Two counties will be reviewed for each state in Wave 1 Processing. A county name in Italics indicates that the county was reviewed during testing and maps do not need to be produced for these counties. During production, only the county verification file and the state summary file needs to be sent to DSSD. The remaining counties will be a complete review with maps, verification files and summary files.

Alaska (FIPS state code 02)
- Wade Hampton Census Area (FIPS county code 270)
- *Denali Borough (068)*

District of Columbia (11)
- *District of Columbia (001)*

Idaho (16)
- Minidoka (068)
- *Washington (087)*

Minnesota (27)
- *Hennepin (053)*
- Watonwan (165)

Montana (30)
- *Glacier (035)*
- Yellowstone (111)

North Dakota (38)
- Morton (059)
- *Mountrail (061)*

South Dakota (46)
- Brown (013)
- *Shannon (113)*

Wisconsin (55)
- *Florence (037)*
- Racine (101)

Wave 2

One county will be reviewed for states that are in Wave 2 of the Address Listing operation. This will be a complete review with maps, verification files and summary files.

State	County
Arkansas (05)	Lee (077)
Arizona (04)	Apache (001)
Colorado (08)	Conejos (021)
Connecticut (09)	Windham (015)
Hawaii (15)	Kauai (007)
Kentucky (21)	Union (225)
Louisiana (22)	Madison (065)
Massachusetts (25)	Suffolk (025)
Maine (23)	Piscataquis (021)
Mississippi (28)	Jefferson (063)
Nebraska (31)	Thurston (173)
New Hampshire (33)	Carroll (003)
New Mexico (35)	Guadalupe (019)
New York (36)	Franklin (033)
Nevada (32)	Humboldt (013)
Oregon (41)	Jefferson (031)
Pennsylvania (42)	Forest (053)
Rhode Island (44)	Bristol (001)
Tennessee (47)	Haywood (075)
Utah (49)	San Juan (037)
Vermont (50)	Essex (009)
Washington (53)	Franklin (021)
West Virginia (54)	McDowell (047)
Wyoming (56)	Carbon (007)

Wave 3

The review for Wave 3 will be in two parts. A complete review will be done of one county in California, Illinois and Texas. Maps, verification files and summary files will be generated for these counties. The remaining states will have one county checked by the SAS program. No maps will be generated for these counties. Only the verification and summary files will be sent to DSSD.

Complete Review

State	County
California (06)	San Fransisco (075)
Illinois (17)	Cook (031)
Texas (48)	El Paso (141)

Note: Cook County, IL is partitioned on TIGER. Pick one of the sub-county partitions for the review.

Computer Program Review Only

State	County
Alabama (01)	Macon (087)
Delaware (10)	Kent (001)
Florida (12)	Gadsden (039)
Georgia (13)	Hancock (141)
Iowa (19)	Muscatine (139)
Indiana (18)	Grant (053)
Kansas (20)	Grant (067)
Maryland (24)	Somerset (039)
Michigan (26)	Saginaw (145)
Missouri (29)	Pemiscot (155)
North Carolina (37)	Warren (185)
New Jersey (34)	Salem (033)
Ohio (39)	Erie (043)
Oklahoma (40)	Adair (001)
South Carolina (45)	Allendale (005)
Virginia (51)	Charles City (036)

Wave 4

The Wave 4 review will be a computer program review only.

State	County
Puerto Rico (72)	Florida Municipio

Appendix H
MAF Housing Unit Status Code

This appendix lists the housing unit status codes. The variable UNITSTAT, as allocated on the MAF, is the field that identifies the housing unit status of an address.

UNITSTAT Legal values:

1 = Valid Living Quarters
2 = Demolished
3 = Open to the elements
4 = Nonexistent
5 = Provisional Add
6 = Under Construction
7 = Duplicate
8 = Vacant Trailer Pad
9 = Burned Out
10 = Boarded Up
11 = Unable to locate
12 = Seasonal
13 = Vacant
15 = Map Spotted Unit with insufficient information
30 = MAF unit moved to another MAF partition
31 = Other uninhabitable

Appendix I
Rules for Housing Unit Count Allocation During PES Clustering

Type of Enumeration Area	Rules/Conditions	Clustering HU count
TEA 1: Mailout/Mailback	1-to-1 or many-to-1 correspondence between 1990 tab block and 2000 collection block	use higher of MAF or 1990 count
	1-to-many correspondence	use MAF count
TEA 2: Update/Leave		use MAF count
TEA 3: List/Enumerate		use 1990 count
TEA 4: Remote Alaska	exclude from PES	
TEA 5: Rural Update/Enumerate		use MAF count
TEA 6: Military	1-to-1 or many-to-1 correspondence between 1990 tab block and 2000 collection block	use higher of MAF or 1990 count
	1-to-many correspondence	use MAF count
TEA 7: Urban Update/Leave	1-to-1 or many-to-1 correspondence between 1990 tab block and 2000 collection block	use higher of MAF or 1990 count
	1-to-many correspondence	use MAF count
TEA 8: Update/Leave to Mailout/Mailback conversion		use MAF count
TEA 9: Mailout/Mailback to Update/Leave conversion		use 1990 count

Appendix J
Equivalency File Layout

The following is the record layout for the Block Equivalency File relating 1990 tabulation blocks to 2000 collection blocks.

Field	Type	Length	Description
ST	CHAR	2	State Code from GTUBAN
RS1		1	Space
COU	CHAR	3	County code from GTUBAN
RS2		1	Space
TRACTBAS	CHAR	4	Tract/Block Numbering area base from GTUBAN
TRACTSUF	CHAR	2	Tract/Block Numbering area suffix from GTUBAN
RS3		1	Space
BLOCKBAS	CHAR	3	Block Base from BKARA
TAB90SUF	CHAR	1	Block Suffix from BKARA
RS4		1	Space
ST2	CHAR	2	State code from COL2000
RS5		1	Space
COU2	CHAR	3	County code from COL2000
RS6		1	Space
COBLKBAS	CHAR	5	2000 Collection Block base from COL2000

UNITED STATES DEPARTMENT OF COMMERCE
Bureau of the Census
Washington, DC 20233-0001

MASTER FILE

May 3, 1999

DSSD CENSUS 2000 PROCEDURES AND OPERATIONS MEMORANDUM SERIES R-9

MEMORANDUM FOR Robert W. Marx
Chief, Geography Division

From: Howard Hogan *Howard Hogan*
Chief, Decennial Statistical Studies Division

Prepared by: Peter P. Davis *PPD*
Sample Design Team,
Decennial Statistical Studies Division

Subject: Amendment to Census 2000 Specifications for Block Cluster Formation–Reissue

This memorandum is being reissued for inclusion in the official *DSSD Census 2000 Procedures and Operations Memorandum Series*. All text in the version of this memorandum distributed on March 2, 1999 remains the same in this version.

Make two changes to the *Census 2000 Specifications for Block Cluster Formation*. They are as follows:

1. MAF Housing Units with a status of 11, unable to locate, have been removed as a valid housing unit code for block clustering. See Section IV., part A., number 1.

2. A one line addition has been made to the block cluster algorithm:

   For each partition, if the calculated cluster target is more than 25 housing units, then set the cluster target equal to 30 and the desired average for the county, usually 30 housing units, set it equal to 40. See section IV., part C., number 1a.

We will reissue the specifications following production.

cc: DSSD Census 2000 Memorandum Series Distribution List
PES Implementation Team
Statistical Design Team Leaders
Sample Design Team
C. Hantman (GEO), R. Ruiz (GEO)
S. Holt (GEO), K. Todd (GEO)

UNITED STATES DEPARTMENT OF COMMERCE
Bureau of the Census
Washington, DC 20233-0001

MASTER FILE

May 3, 1999

DSSD CENSUS 2000 PROCEDURES AND OPERATIONS MEMORANDUM SERIES R-10

MEMORANDUM FOR Robert W. Marx
Chief, Geography Division

From: Howard Hogan *Howard Hogan*
Chief, Decennial Statistical Studies Division

Prepared by: Peter P. Davis *PPD*
Sample Design Team
Decennial Statistical Studies Division

Subject: Accuracy and Coverage Evaluation (ACE) Survey: Second Amendment to Census 2000 Specifications for Block Cluster Formation–Reissue

This memorandum is being reissued for inclusion in the official *DSSD Census 2000 Procedures and Operations Memorandum Series.* All text in the version of this memorandum distributed on March 11, 1999 remains the same in this version.

The Decennial Statistical Studies Division and the Geography Division have reviewed the Block Cluster test data and have encountered four situations which require specification and program changes.

- The "zero block procedure" was not being applied for clusters which do not meet the target size. The specification needs to be clarified and the program revised accordingly.

- The two Alaska counties selected for detailed review are entirely Remote Alaska. Selecting two Remote Alaska counties was not our intention. This occurred because one of the selected counties has recently been converted to entirely Remote Alaska. Hence, we will select a non-Remote Alaska county and exchange it for one of the Remote Alaska counties originally designated for review.

- To verify the "enclosed block procedure," we need a flag on the output file to indicate when an enclosed block has been collapsed. This additional flag needs to be specified and programmed.

- To properly identify four variables on the output file, we need to enlarge the field length allocated to these variables. Therefore, the Verification and the Decennial Systems and Contracts Management Office (DSCMO) Block Cluster File Layout needs to be revised in both the specification and the program for the Block Area, Block Perimeter, Block Cluster Area, and Block Cluster Perimeter variables.

These changes will require production to be restarted.

The following section gives the specific text changes to the block clustering specification. We will reissue the specification following production.

Specification Changes:

Make four changes/additions to the *Census 2000 Specifications for Block Cluster Formation* as follows:

1. In Section IV, part C, paragraph 3c, note the addition in italics.

   Let block A's closest neighbor be block B. A new block cluster, call it AB, is formed by the combination of block A and its neighboring block, block B. If the total number of housing units is greater than 85 percent of the cluster target then proceed to Step 4: Enclosed Blocks. If the total number of housing units is less than 85 percent of the cluster target then find the closest neighbor, block C, and form ABC. Continue to find the closest neighbor and collapse it into the block cluster until the total number of housing units is greater than 85 percent of the cluster target. *If no closest neighbor block C exists, then proceed to Step 4: Enclosed Blocks and then to Step 5: Zero Neighbor Perimeter Search.*

2. In Appendix G, Production Counties for Review, for the state of Alaska, replace Wade Hampton Census Area, (FIPS county code 270) with Anchorage Borough (FIPS county code 020).

3. In Section V, add a new part, part F which should read as follows.

   F. Collapsed Enclosed Block Indicator

   Enclosed blocks that are forced to collapse need to be identified. Enclosed blocks are those blocks having only one neighbor. Collapsing Enclosed Blocks is step 4 in the Algorithm of Block Cluster Formation. Put an indicator on the output file to denote if the enclosed block procedure was implemented on the block. This provides a check that the algorithm was implemented correctly.

Assign the Enclosed Block Indicator at the block level to the output file as follows:

If the block is an enclosed block and it has been forced to collapse, then assign the Collapsed Enclosed Block Indicator a value of '1',
Else assign the Collapsed Enclosed Block Indicator a value of '0.'

Attach this new block-level variable to the end of Appendix A as follows:

Collapsed Enclosed Block Indicator	83:83
0 = Otherwise	
1 = An enclosed block has been forced to collapse	

4. In Appendix A, Verification and DSCMO Block Cluster File Layout, change the file layout to the revised format below.

Variable	Location
State	1:2
County	3:5
Interim Tract (a.k.a. pseudo-tract)	6:11
Block Number	12:16
Blank	17:17
Cluster Number	18:22
Blank	23:23
Cluster Size code	24:24
1 = Clusters with 0 HUs	
2 = Clusters with 1 HUs	
3 = Clusters with 2 HUs	
4 = Clusters with between 3 and 5 HUs	
5 = Clusters with between 6 and 9 HUs	
6 = Clusters with between 10 and 19 HUs	
7 = Clusters with between 20 and 29 HUs	
8 = Clusters with between 30 and 79 HUs	
9 = Clusters with 80 or more HUs	
Blank	25:25
Block Area (Sq. Miles)	26:33
Blank	34:34
Block Perimeter (Miles)	35:40
Blank	41:41
Block Cluster Area (Sq. Miles)	42:49
Blank	50:50
Block Cluster Perimeter (Miles)	51:56
Number of HUs in cluster	57:61

Variable	Location
Number of HUs in block	62:66
Block TEA	67:67
1 = Mailout/Mailback	
2 = Update/Leave	
3 = List/Enumerate	
5 = Rural Update/Enumerate	
6 = Military	
7 = Urban Update/Leave	
8 = Update/Leave to Mailout/Mailback conversions	
9 = Mailout/Mailback to Update/Leave conversions	
TEA Group for Block Cluster	68:68
A = Mailout/Mailback or Urban Update/Leave or Update/Leave to Mailout/Mailback conversions	
B = Update/Leave or Rural Update/Enumerate	
C = List/Enumerate	
D = Military	
E = Mailout/Mailback to Update/Leave conversions	
2000 MAF HUs count	69:73
' ' Blank if no HU count available	
1990 ACF HUs count	74:78
' ' Blank if no HUs count available	
Housing Unit Count Indicator	79:79
1 = from 2000 MAF	
2 = from 1990 ACF	
Invisible Boundary Collapse Indicator	80:80
0 = No	
1 = Yes (Collapsing across Invisible Boundary in BC)	
American Indian Country Indicator	81:81
0 = No American Indian Country	
1 = American Indian Reservation/trust land	
2 = Tribal jurisdiction statistical area/ Alaska Native Village statistical area/ tribal designated statistical area	

Military Indicator	82:82
0 = No Military Area	
1 = Block contains Military Area	
Collapsed Enclosed Block Indicator	83:83
0 = Otherwise	
1 = An enclosed block has been forced to collapse	

cc: PES Implementation Team
Statistical Design Team Leaders
Sample Design Team
C. Hantman (GEO)
R. Ruiz (GEO)
S. Holt (GEO)
K. Todd (GEO)

June 11,1999

DSSD CENSUS 2000 PROCEDURES AND OPERATIONS MEMORANDUM SERIES R-15

MEMORANDUM FOR John Thompson
Associate Director for Decennial Census

and

Preston Jay Waite
Assistant Director for Decennial Census

From: Howard Hogan Howard Hogan
Chief, Decennial Statistical Studies Division

Prepared by: Thomas Mule TM
Sample Design Team
Decennial Statistical Studies Division

Subject: Accuracy and Coverage Evaluation Survey: State Interview Sample Size Estimates

I. INTRODUCTION

This memorandum documents how the Accuracy and Coverage Evaluation (A.C.E.) survey interview sample is expected to be distributed across the 50 states and the District of Columbia. A commitment to a state interview sample size is essential for the Field Division to accurately size staff, space, automation and communication systems, furniture and other equipment required for each of the 12 A.C.E. regional offices (ACEROs) and the Puerto Rico A.C.E. office. Furniture and computer equipment must be ordered now for delivery and installation from mid-summer through early fall of 1999. This is not a final A.C.E. design because the allocation of sample within the states has not yet been determined. Research continues on determining the within-state allocation. We are providing these state level estimates at this time for the planning purposes listed earlier. These numbers are consistent with the approximately 300,000 national housing unit interview sample planned for budgeting purposes in reference 1. These sample size estimates could change due to the budget process or operational resource constraints. These numbers assume that the A.C.E. sample design is contingent on the Integrated Coverage Measurement sample design.

The A.C.E. national sample consists of three components: 1) the general sample, 2) the American Indian Reservation (AIR) sample and 3) the small block cluster sample. The general sample allocation is proportional to total population with a minimum of 1800 housing units in each state and 3750 housing units in Hawaii. The AIR sample allocation is approximately proportional to the 1990 American Indian population on reservations and is described in reference 2. The number of housing units to be interviewed from the small block sample is expected to be low nationally and, consequently, should not significantly impact state interviewing workloads. Therefore, estimates of the interview sample from small blocks are not included in this memorandum. For each state, estimates of the housing unit sample size in terms of interviewing are given for the general sample and AIR sample in Table 1 of the Attachments. The number of interviews is approximately 302,000 for the total U.S. which includes both occupied and vacant housing units.

For completeness, the Puerto Rico sample size is provided in Table 1 as well. The Puerto Rico interview sample size is approximately 15,000 housing units. This is consistent with reference 1.

II. GENERAL SAMPLE ALLOCATION

The general sample is allocated across states proportional to 1998 total population estimates[1] with a minimum of 1800 housing units in each state and 3750 housing units in Hawaii. This is not a final design of the A.C.E. general sample. The Decennial Statistical Studies Division (DSSD) still needs to determine how to allocate the sample within each state. We arrived at this allocation by simulating alternative sample designs and comparing simulated coefficients of variation (CVs) for the 1990 poststrata design. A future memorandum will fully document this research. Features of the allocation are the following:

1. Proportional to State Total Population: Allocating proportional to the state total population is conservative and flexible. A differential sampling plan can be developed in the future using 2000 poststrata information and the results of the A.C.E. initial block cluster listing sample.

2. Minimum of 1800 Housing Units per state: Allocating proportional to total population produced small sample sizes in some states. To address this concern, a number of state minima were examined. The choice of 1800 housing units balanced the gains and losses of the simulated CVs among the poststrata.

3. Sample Size in Hawaii: To support national estimates for Hawaiians and Pacific Islanders as a separate race group requires a larger sample size in Hawaii than 1800 housing units. Alternate sample sizes were examined. The improvement in simulated CVs started to diminish above 3750 housing units.

[1]Ideally, we would have preferred to subtract the American Indian population living on American Indian Reservations from the 1998 state total population estimates to do the general sample allocation. The information to do this was not available. This results in Arizona probably getting a little more sample under this scenario.

III. ACERO SAMPLE SIZES

As noted, the DSSD has not determined how to allocate the sample within each state. There will probably be some differential sampling within California, New Jersey and New York, the three states split by two ACEROs. For the convenience of the Field Division, the DSSD has approximated what the ACERO sample sizes might be. In California, approximately 23,800 housing units (23,600 general sample and 200 AIR sample) will be in the Los Angeles ACERO while approximately 10,040 housing units (9,910 general sample and 130 AIR sample) will be in the Seattle ACERO. In New Jersey, approximately 5,500 housing units (no AIR sample) will be in the New York ACERO while approximately 2,840 housing units (no AIR sample) will be in the Philadelphia ACERO. In New York, approximately 4,910 housing units (4,760 general sample, 150 AIR sample) will be in the Boston ACERO while approximately 13,900 housing units (no AIR sample) will be in the New York ACERO. The ACERO estimates are in Table 2 of the Attachment. Again, these are estimates. Details of the within-state allocation have not been developed.

IV. REFERENCES

1 Memorandum for Hogan from Kostanich, "Accuracy and Coverage Evaluation Survey: Sample Size Estimates," April 30, 1999.

2 Memorandum for Hogan from Kostanich, "Accuracy and Coverage Evaluation Survey: American Indian Reservations Sample Design," April 30, 1999.

cc:
DSSD Census 2000 Procedures and Operations Memorandum Series Distribution List
A.C.E. Implementation Team
Statistical Design Team Leaders
Sample Design Team

Table 1. A.C.E. Interview Sample Allocation

State	Housing Units		
	General	AIR	Total
Alabama	4,470	0	4,470
Alaska	1,800	30	1,830
Arizona	4,800	3,390	8,190
Arkansas	2,610	0	2,610
California	33,510	330	33,840
Colorado	4,080	60	4,140
Connecticut	3,360	0	3,360
Delaware	1,800	0	1,800
District of Columbia	1,800	0	1,800
Florida	15,300	30	15,330
Georgia	7,830	0	7,830
Hawaii	3,750	0	3,750
Idaho	1,800	180	1,980
Illinois	12,360	0	12,360
Indiana	6,060	0	6,060
Iowa	2,940	0	2,940
Kansas	2,700	30	2,730
Kentucky	4,050	0	4,050
Louisiana	4,470	0	4,470
Maine	1,800	30	1,830
Maryland	5,280	0	5,280
Massachusetts	6,300	0	6,300
Michigan	10,080	150	10,230
Minnesota	4,860	300	5,160
Mississippi	2,820	90	2,910
Missouri	5,580	0	5,580
Montana	1,800	720	2,520
Nebraska	1,800	90	1,890
Nevada	1,800	150	1,950
New Hampshire	1,800	0	1,800
New Jersey	8,340	0	8,340
New Mexico	1,800	2,100	3,900
New York	18,660	150	18,810
North Carolina	7,740	120	7,860
North Dakota	1,800	360	2,160
Ohio	11,490	0	11,490
Oklahoma	3,420	240	3,660
Oregon	3,360	90	3,450
Pennsylvania	12,300	0	12,300
Rhode Island	1,800	0	1,800
South Carolina	3,930	0	3,930
South Dakota	1,800	810	2,610
Tennessee	5,580	0	5,580
Texas	20,280	30	20,310
Utah	2,160	210	2,370
Vermont	1,800	0	1,800
Virginia	6,960	0	6,960
Washington	5,850	510	6,360
West Virginia	1,860	0	1,860
Wisconsin	5,370	300	5,670
Wyoming	1,800	150	1,950
U.S. Total	291,510	10,650	302,160
Puerto Rico	15,000	0	15,000

Table 2. A.C.E. Regional Office Interview Estimates

ACERO Code	A.C.E. Regional Office	Housing Units		
		General	AIR	Total
21	Boston	21,620	180	21,800
22	New York	19,400	0	19,400
23	Philadelphia	24,020	0	24,020
24	Detroit	23,430	150	23,580
25	Chicago	23,790	300	24,090
26	Kansas City	22,110	570	22,680
27	Seattle	22,720	940	23,660
28	Charlotte	28,260	120	28,380
29	Atlanta	27,600	30	27,630
30	Dallas	27,570	120	27,690
31	Denver	23,640	8,040	31,680
32	Los Angeles	27,350	200	27,550
	U.S. Total	291,510	10,650	302,160
	Puerto Rico	15,000	0	15,000

UNITED STATES DEPARTMENT OF COMMERCE
Bureau of the Census
Washington, DC 20233-0001

September 22,1999

MASTER FILE

DSSD CENSUS 2000 PROCEDURES AND OPERATIONS MEMORANDUM SERIES R-18

MEMORANDUM FOR Howard Hogan
Chief, Decennial Statistical Studies Division

From: Donna Kostanich DK
Assistant Division Chief, Sampling and Estimation
Decennial Statistical Studies Division

Prepared by: James Farber DF for JF
Sample Design Team

Subject: Accuracy and Coverage Evaluation Survey: Sample Reduction Overview

I. INTRODUCTION

The purpose of this document is to present an overview of the Accuracy and Coverage Evaluation (A.C.E.) sample reduction. The goals of the A.C.E. reduction are three-fold. First, the number of block clusters must be reduced from the listing sample size. The listing sample was selected under the previous 750,000 housing unit design, while the A.C.E. is a 300,000 housing unit design. Second, in reducing the block cluster sample size, the sample sizes for population subgroups that have historically been undercounted in the census should be increased relative to other subgroups to achieve reliable A.C.E. population estimates. The details of differential sampling by demographic group are still in research. The last goal of the A.C.E. reduction is to reduce the variance contribution from block clusters that may potentially be outlier clusters and thus exert undue influence on the population estimates. These clusters are identified by comparing their census and A.C.E. listing housing unit counts. The details of differential sampling for these clusters is also currently in research.

The A.C.E. reduction is one of the processes involved in determining which housing units will be interviewed in the A.C.E. survey. The A.C.E. reduction is a subsample of the medium (3 to 79 housing units) and large (80+ housing units) block clusters previously selected for the A.C.E. listing sample. Following the A.C.E. reduction, the small block cluster subsampling and large block cluster subsampling operations occur. The number of small block clusters (0 to 2 housing units) is reduced during small block cluster subsampling, and the number of housing units in large block clusters is reduced in large block cluster subsampling. The A.C.E. interview sample consists of those housing units in block clusters or segments of block clusters selected for interview in the A.C.E. reduction and subsequent subsampling processes.

II. A.C.E. REDUCTION DESIGN

The following are features of the planned design for the A.C.E. reduction:

- The A.C.E. reduction, along with small block cluster subsampling and large block cluster subsampling, is designed to achieve the national interview housing unit sample size of approximately 300,000 housing units.

- The A.C.E. sample is designed separately by state with the national sample allocated to states proportional to population with a minimum sample size of 1800 housing units.[1]

- Block clusters in Puerto Rico will not be subsampled in the A.C.E. reduction. All Puerto Rico clusters will be retained in the A.C.E. interview sample.

- Block clusters on American Indian Reservations[2] will not be subsampled in the A.C.E. reduction. All American Indian Reservation clusters will be retained in the A.C.E. interview sample.

  Clusters on Tribal Jurisdiction Statistical Areas, Tribal Designated Statistical Areas, and Alaska Native Village Statistical Areas probably will be subsampled in the A.C.E. reduction. This issue is currently unresolved.

- Small block clusters will not be subsampled in the A.C.E. reduction. A separate subsampling operation will reduce the number of small block clusters in the A.C.E. interview sample.

---

[1] Mule (June, 1999), "Accuracy and Coverage Evaluation Survey: State Interviewing Sample Size Estimates," DSSD Census 2000 Procedures and Operations Memorandum Series R.

[2] The American Indian Reservations includes the associated Trustlands.

- Only medium and large block clusters that are not on an American Indian Reservation and not in Puerto Rico are subsampled in the A.C.E. reduction.

- The calculation of reduction sampling rates is based on the most recent measure of size, the preliminary A.C.E. independent listing housing unit count. These housing unit counts are preliminary because the number is simply a clerical tally of the number of housing units listed in the independent listing book.

- Block clusters that were in the medium stratum at the time of listing sample selection but have 80 or more housing units based on the preliminary listing housing unit count will likely be retained at higher rates to control their weights. In the listing sample, medium clusters were sampled at lower rates than large clusters since large clusters eventually undergo large block subsampling, an operation that increases weights.

- Excluding medium clusters that have 80 or more housing units on the independent list, medium and large block clusters will be subsampled in the A.C.E. reduction at the same relative rates used in listing sample selection. That is, the differential allocation of medium and large clusters in the listing sample will be retained in the reduced sample.

III. RESEARCH ITEMS

The following issues for the A.C.E. reduction are still in research:

- Differential subsampling rates may be used for certain demographic groups, such as minority/non-minority, in states where the population is estimated to be sufficiently heterogeneous and where the listing sample size is sufficiently large. Differential sampling as opposed to proportional sampling could provide more reliable A.C.E. estimates for demographic groups that have historically been undercounted. Research is ongoing to determine whether differential sampling by demographic group is expected to provide variance reduction, and if so, what the differential subsampling rates should be to maximize variance reduction while also controlling weight variation. It is expected that no more than two demographic strata would be formed in a single state to control weight variation.

- Differential subsampling rates may also be used for clusters where the current census housing unit count differs significantly from the A.C.E. independent listing housing unit count. Clusters with significant differences are called "Inconsistent" while other clusters are "Consistent." It is expected that only two strata will be formed, although it is possible that the Inconsistent stratum might be split into two strata depending on the results of research. The definition of a significant difference is unresolved at this point. Possibilities include measures based on absolute or percent differences in the two housing unit counts. Inconsistent clusters are more likely to experience coverage problems and thus should be retained in the A.C.E. interview sample at a higher rate than Consistent clusters. The extent to which Inconsistent block clusters might be differentially sampled is currently unknown. All List Enumerate clusters will likely be considered inconsistent since the census housing unit count in these clusters is unknown at the time of the A.C.E. reduction.

- In many states, it is possible that differential sampling will be used based on both housing unit count differences and demographic groups, and thus these two types of strata need to be integrated in the A.C.E. reduction. To control weight variation, the current plan is to combine these two types of strata into three A.C.E. reduction strata:

    - Minority Block Clusters
    - Non-Minority Inconsistent Clusters
    - All Remaining Clusters

    Alternative combinations are also under consideration. To further reduce weight variation, the first two A.C.E. reduction strata may possibly be given the same sampling rate. Other alternatives to control weight variation are also being researched.

- The plans are to use differential sampling conservatively. The research may show large variance gains by allowing extensive weight variation; these gains may not be achieved for 2000. From the demographic groups perspective, population shifts have probably occurred since 1990 which could cause wide weight variation within demographic groups that we would like to avoid. For the consistency of housing unit counts, the reduction of variance may not be as significant if the targetted extended search program is successful. Further, the relationship of the two housing unit counts is a proxy variable. Even when the two counts are relatively comparable for a cluster, it is possible for there to be coverage problems in the cluster.

cc: Census 2000 Procedures and Operations Memorandum Series Distribution List
Statistical Design Team Leaders
Sampling and Estimation Staff

UNITED STATES DEPARTMENT OF COMMERCE
Bureau of the Census
Washington, DC 20233-0001

MASTER FILE

October 22, 1999

CENSUS 2000 PROCEDURES AND OPERATION MEMORANDUM SERIES R-20

MEMORANDUM FOR Maureen P. Lynch
Assistant Division Chief, ICM Processing
Decennial Statistical Studies Division

From: Donna Kostanich FN for DK
Assistant Division Chief, Sampling and Estimation
Decennial Statistical Studies Division

Prepared by: Douglas Olson DO

Subject: Accuracy and Coverage Evaluation Survey – Identification and Sampling of Block Clusters for Targeted Extended Search

I. Introduction

In 1990, every block cluster sampled in the Post Enumeration Survey (PES) was also searched for persons in the surrounding blocks in an operation called "Surrounding Block Search". Experience with that operation demonstrated that searching around every cluster was not necessary because most clusters did not show any errors in the original P- or E- samples related to geography. What errors there were tended to occur in bunches, in which large groups of housing units were incorrectly enumerated or matched because of an error in locating them geographically. Because such errors tend to be clustered, it was decided for the 2000 Accuracy and Coverage Evaluation (A.C.E.) that only 20 percent of clusters would have their surrounding areas searched, and that the clusters would be selected based on a criterion that will result in the sample inclusion of a high number of clusters with large numbers of census geocoding errors. Such a search is called a "Targeted Extended Search" or "TES."

In addition to the deliberate selection of "bad" clusters, the search will be targeted by address. Before the A.C.E. person interviewing (during which TES will be performed concurrently) the Census Bureau will conduct Housing Unit Matching. This process will identify what housing units represent the Geocode Errors or Address Non-matches that flag the need to search in the surrounding area. Unlike in 1990 Surrounding Block Search, the 2000 TES will search only for the addresses that are flagged as Geocode Errors or Address Non-matches, which will make them candidates for TES. The persons living in such housing units will be considered as "TES persons" and weighted in the P- and E-samples to reflect the probability that their block cluster had been chosen for TES.

These specifications describe the steps necessary to select the TES clusters and assign them TES weights. Those data, along with additional cluster information needed to check the validity of the sample selection, will be included in the Sample Design File.

II. Overview of Process

In outline, there are four basic steps to selecting the TES clusters:

- Determine sampling parameters and create a TES Parameter File
- Select TES Clusters
- Update the Sample Design File with the results of TES sample selection
- Verify the sample selection

Because some blocks are going to generate more P-sample and E-sample persons through TES than others, it would be desirable to design a sample in which clusters have different probabilities of being included in TES. Before the TES sample has to be selected, there will be certain information available about the clusters that will be helpful in targeting the selection of clusters for TES. Specifically, we would like to be certain to include in TES all clusters which:

- Were re-listed
- Have many housing units coded as address non-match, unresolved address match or geocode error in housing unit matching
- Have a high weighted total of housing units in these categories

Therefore, the TES clusters will be of two types:

- Those included in the TES with certainty because they show the characteristics above
- Clusters selected randomly from those that do *not* show these characteristics

The TES interview workload will include

- All re-listed clusters,
- The 5 percent of clusters with the most census geocode errors and A.C.E. address nonmatches
- The 5 percent of clusters with the most weighted census geocode errors and address nonmatches
- 10 percent of the remaining 90 percent of clusters (excluding relisted and list/enumerate clusters) selected using a systematic sample

At the end of the selection, all clusters will be assigned a variable TESFLAG that identifies their TES status and a variable TETES, a weight that reflects the probability that the cluster be included in TES. There are a few types of clusters that require special handling:

- The Puerto Rico A.C.E. operations are similar but independent of the United States. The

TES selection for the Puerto Rico A.C.E. operations will be implemented independently.

- List/Enumerate clusters will not be included in this part of TES selection. L/E clusters are out of scope for sample selection purposes. TES special procedures for these areas are currently under development.

II. Detailed Sampling Procedures

The TES sample selection is a national-level process, one of the few "one-shot" A.C.E. related operations. It will be implemented completely at once. All preliminary operations, listed under "Assumptions" below, must be completed before the TES clusters are selected.

Assumptions

Before the beginning of the TES sample selection, the following will be complete and available:

- A.C.E. sample selection and A.C.E. small block subsampling.
- A.C.E. sample reduction. The above sampling activities will be reflected on the Sample Design File.
- All initial housing unit matching operations.
- Identification of clusters to be relisted.

Large-block subsampling will *not* be completed in time for TES sample selection.

Data Sources

1. Input Files:
    a. HUMARCS_ACCT2K (MaRCs housing unit account file)-- This is a block cluster level file which includes one record for each cluster in the A.C.E. sample. This file will include the results of all the above sampling activities, as well as the initial housing unit matching results. For example, for each block cluster it will have a count of independent listing addresses not matched to census addresses that were confirmed to exist in the A.C.E. block cluster.
    b. ACE2000_SDF (Sample Design File) – include all listed clusters, whether selected for A.C.E. or not. Before starting TES, it will include the final weight of all clusters in the A.C.E. sample, including small block subsampling weights, and additional sampling related codes

2. Created during processing:
    a. TESPARAM (TES Parameter File) – Includes only two records with several variables to be used in TES sample selection.
    b. TESCLUST – file of all A.C.E. sample clusters, which will include variables and information required for TES sample selection.

3. Output File:
    ACE2000_SDF?.<mmddyy> (The Sample Design File, The "?" refers to the version number of the Sample Design File, which will be updated on a flow basis. "<mmddyy>" is the date on which the most recent version of the Sample Design File was created.)-- The TES sampling operation will generate a subsequent version of the Sample Design File, with an updated version number and date. The data used in selecting the TES sample, and the sampling output itself will be included in this file for subsequent use during several production operations.

Operations

A. Create TES Parameter File (TESPARAM)

File TESPARAM (see file layout in the Attachment) contains parameters that will be used to select the TES sample. It will have two records, one for the U.S. and one for Puerto Rico. The first record has PRFLAG set equal to 0 to indicate that this record is for the U.S. The second record is for Puerto Rico. Set PRFLAG equal to 1. This record holds the sampling parameters for Puerto Rico. The other fields that need to be set before beginning the process are:

- TESRATE - The fraction (expressed as a decimal) of clusters that will be included in the TES sample. The 20 percent sampling fraction **does not** include relisted clusters. Set TESRATE equal to .20.

- UNWCRATE - The fraction (expressed as a decimal) of clusters that will be used for TES selected with certainty based on the *unweighted* number of interesting housing units (to be defined later). Set UNWCRATE to .05.

- WGTCRATE- The fraction (expressed as a decimal) of clusters that will be used for TES selected with certainty based on the *weighted* number of interesting housing units (to be defined later). Set WGTCRATE to.05.

- SUMORDIF – Flag to indicate wether TES will be based on the sum or difference of the interesting housing units in the P- and E-samples. Flag equals 1 if using the sum, 0 if using the difference.

- The remaining fields (NUMCLUST, RELISTCT, LECOUNT, UNCERNUM, WTCERNUM, SAMPSIZE and TETES) will be updated later and initially must be set to zero.

B. Create TES Cluster File (TESCLUST).

This file will contain one record for every A.C.E. sample cluster and includes the information needed to perform TES sample selection. This file will include all the records and a subset of data fields from the HUMaRCS Account File plus additional variables from the Sample Design File.

1. For each record in the Account file copy to file TESCLUST the fields:

   CLUST - cluster number
   CURCI - housing units with match code "CI"or confirmed address non-match
   CURUI - housing units with match code "UI" or unconfirmed address match
   CURGE - housing units with match code "GE" or census geocode error
   RELIST - relist flag = 1 if cluster re-listed, 0 otherwise
   STATE - 2-digit FIPS state code
   CMDONE - Computer match done code

2. From the Sample Design File, add the following fields into TESCLUST, using the same variable names:

   WEIGHTC - A.C.E. cluster weight
   SS - A.C.E. sampling stratum
   ARST - A.C.E. sample reduction stratum
   SBCSS - small block cluster sampling stratum

3. Create additional fields that will be used in the TES selection, and assign initial values:

   Variables SAMPSTRT and PRFLAG will be used to identify sampling stratum and Puerto Rico/United States sampling process, respectively.

   SAMPSTRT, concatenate fields STATE, SS, ARST, SBCSS
   PRFLAG=1 if STATE=72, and PRFLAG=0 otherwise

   We want to target for certainty inclusion in TES clusters that have many address non-matches, unresolved address matches and census geocode errors. This information was extracted from the Housing Unit Account file in fields CURCI, CURUI and CURGE. We will need to know the total number of such housing units, and both a weighted and unweighted basis. The criterion that will be used to select the sample is a function of these counts. The exact form of the function is yet to be determined. Therefore, we have to create several variables to hold the weighted and unweighted sum and difference of these totals, and one variable to use as a sort variable once the sample selection criterion is agreed upon. Compute the following variables:

   SUMUNIHU = CURCI + CURUI + CURGE
   DIFUNIHU = Absolute value of (CURCI + CURUI- CURGE)
   SUMWTIHU = WEIGHTC * SUMUNIHU, rounded to nearest integer
   DIFWTIHU = WEIGHTC * DIFUNIHU, rounded to nearest integer
   SRTUNIHU = SUMORDIF*SUMUNIHU + (1 - SUMORDIF)*DIFUNIHU
   SRTWTIHU = SUMORDIF*SUMWTIHU + (1 - SUMORDIF)*DIFWTIHU

A few variables will ultimately be copied to the Sample Design File to identify the TES selection. For the time being, they need to be put into TESCLUST and initialized as follows:

TESELECT="Z"
TESFLAG=0
TETES=0
TESN=0
RSTES=0

4. Get a count of the number of clusters in the U.S. and Puerto Rico. Count the total number of records in TESCLUST for PRFLAG=0 and PRFLAG=1 separately. Put the total in field NUMCLUST in file TESPARAM. The first record will show the number of in-sample A.C.E. clusters in the U.S. and the second will show the number of in-sample A.C.E. clusters in Puerto Rico.

C. Identify the List/Enumerate clusters

List/Enumerate clusters will be excluded from the TES sampling operations. For these clusters, which can be identified by variable CMDONE=5, update the selection variables to reflect that they will not be part of TES:

TESELECT="O", cluster is out-of-scope for TES
TESFLAG=2, cluster is not eligible for TES
TETES=<blank>, TES Weight not relevant for these clusters
TESN=0, no order assignment used
RSTES=<blank>, random start not relevant for these clusters

D. Identify relisted clusters

All relisted clusters, identifiable by variable RELIST=1, will be included in TES. For each record in TESCLUST, if RELIST=1, set the following variables:

TESELECT="R", cluster was relisted
TESFLAG=1, cluster will be included in TES
TETES=1, the cluster's TES Weight will equal one
TESN=0, no order assignment used
RSTES=0, random start not used

E. Selection of Certainty Clusters

There are two phases to selecting the certainty cases for TES--weighted and unweighted. The clusters (records) with the highest totals of SRTUNIHU and SRTWTIHU will be selected with certainty for inclusion in TES.

1. Since Re-list and List/Enumerate clusters are not of interest for the rest of the process, get counts of both types of assignments and put them into the TES Parameter file. For both PRFLAG=0 and PRFLAG=1, count the number of TESELECT="R" records and put the total in the TESPARAM field RELISTCT . Count the number of TESELECT="O" clusters and put the total in TESPARAM field LECOUNT.

2. Calculate the number of weighted and unweighted certainty cases needed. Using the TESPARAM file, calculate separately for PRFLAG=0 and =1:

UNCERNUM = (NUMCLUST - LECOUNT) x UNWCRATE, rounded to nearest integer.
WTCERNUM = (NUMCLUST - LECOUNT) x WGTCRATE, rounded to nearest integer.
If either of the variables UNCERNUM or WTCERNUM is negative, set to zero.

3. First, we'll select the certainty cases based on unweighted criteria. Sort TESCLUST by the variables TESELECT, PRFLAG, SRTUNIHU, WEIGHTC, CLUST from largest value to smallest value.

4. The clusters are in the order in which we want them to select the unweighted certainty cases. Only clusters that currently have TESELECT="Z" are eligible, and we need separate draws for PRFLAG=0 and PRFLAG=1. So, within the group TESELECT="Z" and PRFLAG=0, and again within TESELECT="Z" and PRFLAG=1, select the first UNCERNUM records for inclusion with certainty. For those records set the fields:

   TESELECT="U", cluster was selected with certainty based on unweighted criteria
   TESFLAG=1, cluster will be included in TES
   TETES=1, the cluster's TES Weight will equal one
   RSTES=0, randomization not used in selecting cluster
   TESN=0, cluster order not used

5. Now select the certainty cases based on weighted criteria. Sort TESCLUST by the variables TESELECT, PRFLAG, SRTWTIHU, CLUST from largest value to smallest value.

6. For both PRFLAG=0 and PRFLAG=1, select the first WTCERNUM records for which TESELECT="Z" in order by the sort above. These have been selected for inclusion in TES, based on their weighted count of interesting housing units. For those records set the fields:

   TESELECT="W", cluster was selected with certainty based on weighted criteria
   TESFLAG=1, cluster will be included in TES
   TETES=1, the cluster's TES Weight will equal one
   RSTES=0, randomization not used in selecting cluster

TESN=0, cluster order not used

F. Selection of the TES sample

We want to draw a systematic sample of an appropriate percentage of the size of the original cluster universe (excluding List/Enumerate clusters.) At this writing, we think that percentage will be 10 percent, but need to design in some flexibility in case that has to be changed, so use the values in the TES Parameter file to calculate the sample size separately for PRFLAG=0 and =1, and put the result into the TES Parameter File:

SAMPSIZE = [TESRATE x (NUMCLUST - LECOUNT)] - UNCERNUM - WTCERNUM, rounded to the nearest integer.

To get a sample of the desired size, we need a take-every that produces a sample of the correct size. Calculate the take-every and put into the TES Parameter file:

TETES =
(NUMCLUST - LECOUNT - RELISTCT - UNCERNUM - WTCERNUM) / SAMPSIZE , rounded to six decimal places.

Samples will be taken separately for each PRFLAG:

1. Sort the block clusters within each PRFLAG by TESELECT, SAMPSTRT, CLUST. All subsequent operations will be performed on clusters where TESELECT="Z".

2. Number the block clusters from 1 to N, where N is the number of block clusters with the appropriate PRFLAG. Put these indexes to variable TESN.

3. get the take-every (TETES) from the TESPARAM file.

4. Generate a sequence of numbers $TESRAND_1$, ..., $TESRAND_n$ as follows:

   a. generate a random number (RN) between 0 and 1 with 10 decimal places.

   b. Calculate a random start, RSTES, which equals RN×TETES. Round this number to six decimal places. Put into field RSTES for every cluster in the sampling universe (i.e. all clusters where TESELECT="Z", for the appropriate PRFLAG.)

   c. Let $TESRAND_1$ = RSTES.

   d. Calculate $TESRAND_J = TESRAND_{J-1}$ + TAKEVERY for J = 2, 3, ...,n, where n is the largest integer such that [RSTES + (n - 1)×TAKEVERY] ≤

N.

e. Round each $TESRAND_J$ up to the nearest integer (an integer rounds to itself).

6. Each cluster with TESN equal to one of the rounded values of $TESRAND_J$, J = 1, 2, ..., n, is in the TES sample. Set the following variables:

   TESELECT="S"
   TESFLAG=1
   TETES=TAKEVERY

7. Each cluster with TESN **not** equal to one of the rounded values of $TESRAND_J$, J = 1, 2, ..., n, is not in the sample. For these clusters set:

   TESELECT="N"
   TESFLAG=0
   TETES=0

G. All clusters have now been selected into or out of TES. Copy the variables listed on page 14 into the Sample Design File.

IV . Verification

Files TESPARAM, TESCLUST and the Sample Design File will be used for verification.

A. Verify that all relisted clusters (RELIST=1 in TESCLUST) are in the Sample Design File with TESELECT="R" and TETES = 1.

B. Verify that all List/Enumerate clusters (CMDONE=5 in the Sample Design File) are in the Sample Design File with TESELECT="O" and TETES = 0.

C. Of those clusters where TESELECT="U", identify the minimum value of CURCI+CURUI+CURGE (or difference if used in sampling). Check that all clusters whose total is greater than that value are TESELECT="U" and that clusters whose total is smaller than that are not TESELECT="U". Ignore RELIST clusters in this step. Check that the number TESELECT="U" is UNCERNUM.

D. Of those clusters where TESELECT="W", identify the minimum value of CURCI+CURUI+CURGE (or difference if used in sampling), multiplied by the cluster weight. Check that all clusters whose total is greater than that value are TESELECT="W" and that clusters whose total is smaller than that are not TESELECT="W". Ignore RELIST and TESELECT="U" clusters in this step. Check that the number of TESELECT="W" is WTCERNUM.

E. Considering only clusters whose TESELECT="S" or TESELECT="N":

1. Sort by SAMPSTRT, CLUST
2. Check that the random starts are in [0, TAKEVERY)
3. Check that indexes were assigned correctly, (TESN increment by 1 for TESELECT="S" or TESELECT="N")
4. Duplicate the selection of TESRANDj using RSTES and TAKEVERY
5. Check that the number of TESELECT="N" is about eight times TESELECT="S"

V. Testing

Since the TES selection will be performed only once for the whole country, we would like to perform a dry run before the actual sample has to be selected. The Variance estimation staff will furnish to the Coverage Measurement Processing staff a set of test files corresponding in layout to those needed for the 2000 TES:

DRPARAM (equivalent to TESPARAM)
A sample HuMARCS Account file
DRSDF1 (equivalent to ACE2000_SDF)

These files will include all the variables used for 2000 TES selection in their proper fields. For purposes of the dry run, we may change some state code to STATE=72 to simulate the effect of Puerto Rico. The output expected will be an updated file DRSDF2, corresponding in layout to the 2000 Sample Design File after TES sampling is completed.

As output, we would like to receive all the output files from the TES selection:

DRPARAM (updated during processing)
TESCLUST (created during processing)
DRSDF2 (an updated version of DRSDF1, reflecting the TES selection)

Since we intend to design the files with exactly the same variable names and layout, the program should be exactly the same as the final TES program, except for the file names. The files will be delivered for testing not later than January 14, 2000 to be completed by the software testing deadline of February 15, 2000.

The target date to complete all software development and testing is February 15, 2000. That is, the TES sample selection computer system will be ready for production on 02-15- 00.

Attachment

File Layouts

File: HUMARCS_ACCT2K (MaRCs housing unit account file, one record per cluster in A.C.E.)
Fields used for TES:

Field	Description	Width	Fields	
CLUST	Cluster Number	6	1-	6
STATE	State Code	2	424-	425
CURCI	Current HU's with Match="CI"	5	320-	324
CURUI	Current HU's with Match="UI"	5	315-	319
CURGE	Current HU's with Match="GE"	5	370-	374
RELIST	Relist Flag (0=No, 1=Yes)	1	410-	410

File: ACE2000_SDFV?.mmddyy (Sample Design File, one record per listed cluster)
Fields input for use in TES:

Field	Description	Width	Fields	
CLUST	Cluster Number	5	21-	25
WEIGHTC	Unbiased weight for A.C.E. cluster	12	334-	345
SS	Sampling Stratum	1	55-	55
ARST	A.C.E. reduction stratum	2	190-	191
SBCSS	Small block cluster sampling stratum	2	306-	307

File TESCLUST (one record for each cluster in A.C.E.)

Field	Description	Field Width	Source/Initial Value
CLUST	Cluster number	6	Acct File
CURCI	Current HU's with Match="CI"	5	Acct File
CURUI	Current HU's with Match="UI"	5	Acct File
CURGE	Current HU's with Match="CI"	5	Acct File
RELIST	Relist Flag (0=No, 1=Yes)	1	Acct File
STATE	State Code	2	Acct File
CMDONE	Computer Match Done Code	1	Acct File
WEIGHTC	A.C.E. Sampling Weight	12.6	SD File
SS	Sampling Stratum	1	SD File
ARST	A.C.E. Reduction Stratum	2	SD File
SBCSS	Small Cluster Subsampling Start.	2	SD File
SAMPSTRT	Sampling Stratum for TES	7	STATE|| SS || ARST || SBCSS
PRFLAG	Puerto Rico Flag	1	=1 if STATE=72, 0 otherwise
SUMUNIHU	Sum Unweighted Interesting HU's	5	CURCI + CURUI + CURGE
DIFUNIHU	Diff. Of Unwgt. Interesting HU's	5	| CURCI + CURUI - CURGE |
SUMWTIHU	Sum of Weighted Interesting HU's	5	WEIGHTC x SUMUNIHU
DIFWTIHU	Diff. of Weighted Interesting HU's	5	WEIGHTC x DIFUNIHU
SRTUNIHU	Sort for Unwgt. Interesting HU's	5	SUMUNIHU or DIFUNIHU
SRTWTIHU	Sort of Weighted Interesting HU's	5	SUMWTIHU or DIFWTIHU
TESELECT	TES Selection Type	1	"Z"
TESFLAG	TES Selected Flag	1	0
TETES	TES Take-every	12.6	0
RSTES	Random Start used in sampling	12.6	0
TESN	Index value used in sampling	6	0

The Sample Design File (ACE2000_SDF?.mmddyy) one record per listed cluster
This file has many fields, most unrelated to TES. In addition to other fields, the following have to be added for TES, all will be copied from TESCLUST after selection is finished:

Field	Description	Field Width	Fields	
CURCI	Current HU's with Match="CI"	5	676-	680
CURUI	Current HU's with Match="UI"	5	682-	686
CURGE	Current HU's with Match="CI"	5	688-	692
TESELECT	TES Selection Type	1	694	
TESFLAG	TES Selected Flag	1	696	
RSTES	TES Random Start	12.6	698-	709
TETES	TES Take-every	12.6	710-	721
TESN	Index value used in sampling	6	722-	727

The possible values for TESELECT in the TESCLUST and Sample Design Files:

Code	Description	Prob of Selection	TETES
Z	Initial value; should not be present after selection is completed		
R	Re-listed cluster, must be included in TES	100 percent	1
U	Certainty selection based on unweighted criterion	100 percent	1
W	Certainty selection based on weighted criterion	100 percent	1
S	Selected by 1-in-9 sampling of non-certainty cases	11 percent	9
N	Not selected for TES	89 percent	0
O	Out of Scope for TES	0 percent	<blank>

The TES Parameter File: TESPARAM (two records, containing global variables)
Input before beginning of processing:

Field	Description	Width	Initial Value
PRFLAG	Equals 1 for Puerto Rico, 0 otherwise	1	0/1
TESRATE	Overall TES selection rate	8.6	.20
UNWCRATE	Portion selected with certainty, based on unweighted count	8.6	.05
WGTCRATE	Portion selected with certainty, based on weighted count	8.6	.05
SUMORDIFF	Flag for selection based on sum or difference (1=Sum, 0=Diff)	1	0 or 1
Created during processing:			
RELISTCT	Count of re-listed clusters	5	0
LECOUNT	Count of List/Enumerate clusters	5	0
UNCERNUM	Certainty cases based on unweighted	5	0
WTCERNUM	Certainty cases based on weighted	5	0

NUMCLUST	Number of clusters from which TES drawn	5	0
SAMPSIZE	The desired number of sampled cases	5	0
TAKEVERY	Take-Every used in sampling	12.6	0

UNITED STATES DEPARTMENT OF COMMERCE
Bureau of the Census
Washington, DC 20233-0001

December 16, 1999

MASTER FILE

DSSD CENSUS 2000 PROCEDURES AND OPERATIONS MEMORANDUM SERIES R- 22

MEMORANDUM FOR Howard Hogan
Chief, Decennial Statistical Studies Division

From: Donna Kostanich DK
Assistant Division Chief, Sampling and Estimation
Decennial Statistical Studies Division

Prepared by: Deborah Fenstermaker DF
Sampling Staff

Subject: Accuracy and Coverage Evaluation Survey: Cluster Reduction Contingency Plan

The delay in finishing the delivery of the Master Address File (MAF) Extract files from December 15, 1999 to January 7, 2000 affects plans for implementing the Accuracy and Coverage Evaluation (A.C.E.) cluster reduction. The Decennial Statistical Studies Division (DSSD) staff worked with staff from the Decennial Systems and Contracts Management Office (DSCMO) to develop a plan for dealing with the delay. Attached is our contingency plan.

cc: DSSD Census 2000 Procedures and Operations Memorandum Series Distribution List
Statistical Design Team Leaders Distribution List
Sampling and Estimation Staff

Attachment

A.C.E. Cluster Reduction: Contingency Plans due to Delay in MAF Extract Delivery

This document outlines plans developed in conjunction with the DSCMO staff for implementing the A.C.E. cluster reduction given the delay in the finish of the MAF Extract delivery to the DSCMO from December 15, 1999 to January 7, 2000. These delayed MAF Extracts include updates from the September and November Delivery Sequence File (DSF) processing for Mailout/Mailback areas. The DSSD wants to use these census housing unit counts for the A.C.E. cluster reduction since they will be the most up-to-date counts at the time of the reduction. However, since the delay in delivering these files significantly cuts into the time for implementing and verifying the reduction, putting the successful completion of these activities at risk, we have developed a contingency plan.

BACKGROUND

A major component of the A.C.E. cluster reduction design is stratifying the clusters based on the relationship of current housing unit counts from the A.C.E. independent listing and the census address list. Clusters will be differentially sampled in order to reduce the variance contribution of clusters with significant differences between the census and the independent list housing unit counts. Clusters with significant differences are likely to have high erroneous enumerations and high nonmatch rates. The objective of differentially sampling these types of clusters is to reduce the magnitude of the weights associated with clusters having potentially high coverage measurement implications. It's important to have the most up-to-date housing unit counts in order to stratify the clusters effectively. Misclassification of clusters leads to undesired differential weights.

Under the original plans for the A.C.E. cluster reduction, the most up-to-date census housing unit counts were scheduled to be available on the Decennial Master Address File (DMAF) by December 23, 1999. This timing relied on the MAF Extracts, which contain the September and November DSF updates, being delivered to the DSCMO by December 15, 1999. The independent listing counts are scheduled to be available from the Technologies Management Office (TMO) on December 16, 1999 for all sample clusters. This is a one-time delivery for the entire national sample and no updates will be made after this delivery. The A.C.E. reduction is scheduled to begin the next working day following the DMAF updates, December 27, 1999. The verification and approval of the A.C.E. reduction must be completed by January 21, 2000. This finish date cannot be delayed because too many important activities rely on the timely completion of the A.C.E. reduction, and there is no slack in the schedule.

We have learned that the delivery of the MAF Extracts is delayed until January 7, 2000. The MAF Extracts will be delivered to the DSCMO on a flow-basis, state-by-state. It is our understanding that January 7, 2000 is the date on which the last state file will be delivered. The DSCMO requires about a week to update the DMAF depending on how early the flow of states begins, the magnitude of the updates, and whether there are any reissues of MAF Extract files.

CONTINGENCY PROCESSING PLAN

Here are some highlights and implications of our strategy for dealing with the delay in the MAF Extracts and meeting the planned date of January 21, 2000 for completing the verification of the A.C.E. reduction.

- Implement the reduction as soon as the DMAF is updated with the January MAF Extracts.

- Successful implementation and verification relies on the process working smoothly.
  - → The MAF Extracts have to be delivered on time, and there cannot be any surprises when updating the DMAF or implementing the reduction.

- As a contingency plan, the A.C.E. sample reduction will also be implemented in December, 1999. The December reduction will use the DMAF counts without the January updates.
  - → The December production will not incorporate the September and November DSF updates.
  - → The census counts will only include updates since the July and August deliveries due to the delivery of Puerto Rico, the November updates for Update Leave areas, and any housing units located at GQs from the December updates.

- The DSCMO and the DSSD will attempt to start the December production and verification earlier than originally planned, possibly as early as December 21, 1999, if other priorities allow.
  - → We may be able to run earlier than originally scheduled because the December updating of the DMAF is not as extensive since the DSF updates are not being processed at this time.
  - → We do not intend to update the MAS. We will try our best to run early, but circumstances may prevent this.
  - → Our goal is to completely verify the December results before January 7, 2000.

- Every effort will be made to successfully verify the January production data by January 21, 2000. If by close of business on January 21, 2000, the January data is not verified, then the December data will be the official A.C.E. reduced sample.

- Using the December data as the official results has the following implications:
  - → There may be a potential increase of the variance by 1) introducing undesired differential sampling caused by misclassifying clusters, and 2) losing the ability to

detect clusters with significantly different housing unit counts between the A.C.E. and the census without having the September and November DSF updates.
  - → This is a fallback plan to keep the A.C.E. program on schedule.
  - → The reduction will be done using a different version of the DMAF than that used for the housing unit matching operation.
  - → An extra task is necessary to provide the DMAF housing unit counts from the January update for small block cluster subsampling.

- This two-step process more than doubles the efforts to implement and verify the A.C.E. reduction. Additional efforts are required to manage and organize the system to ensure there is no corruption of files between the December and January processing.
  - → We plan to process in separate computer subdirectories.

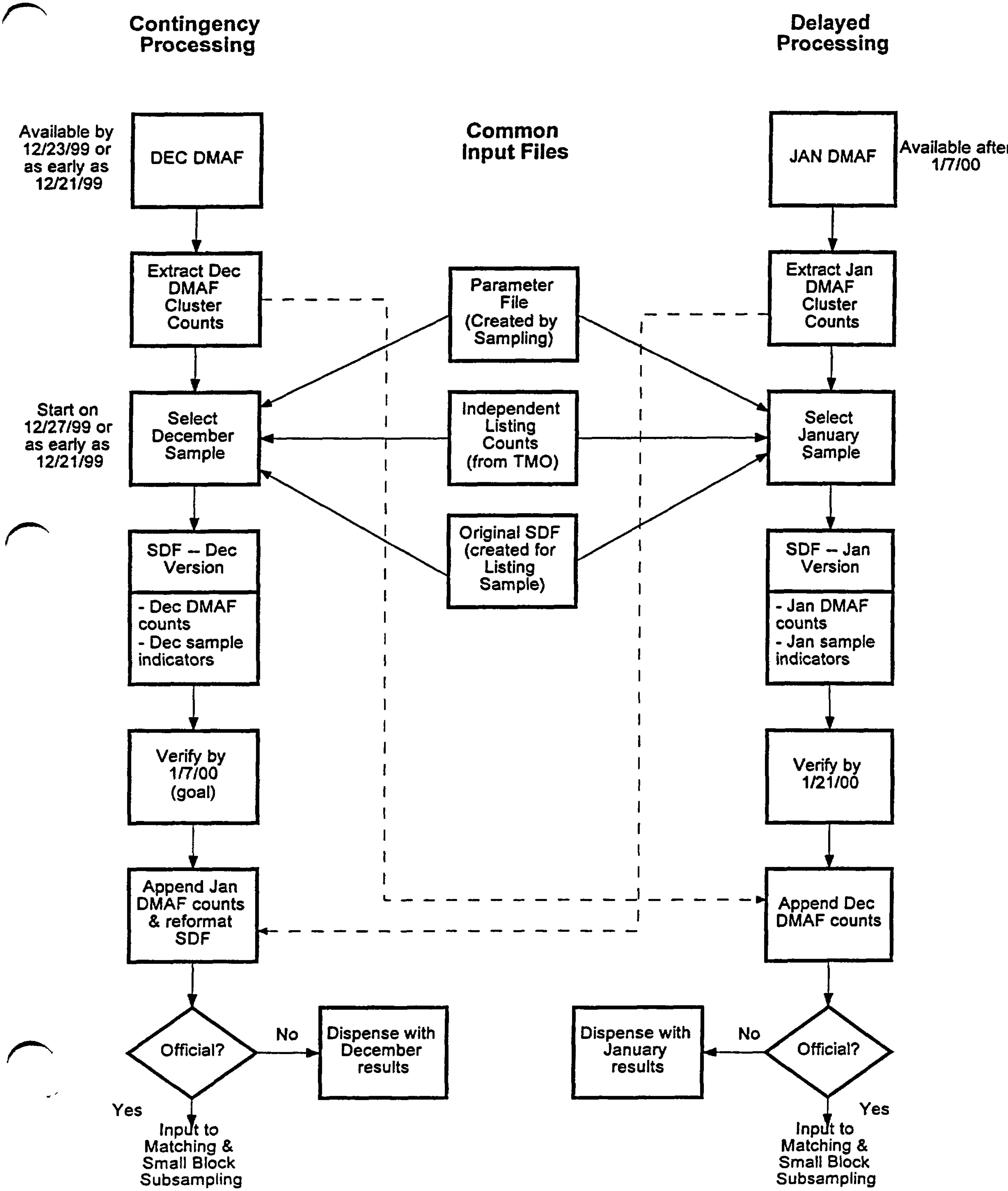
- We need to develop a policy for dispensing of the results which do not become "official". Do we delete these files? If we don't complete the verification of the January production by January 21, do we suspend the work forever and delete the results?

- If the MAF Extracts run late and all states are not delivered to the DSCMO by January 7, we will NOT implement the A.C.E. reduction using an updated DMAF. There is no chance of successfully updating the DMAF, implementing the reduction, and verifying the results by January 21, 2000.
  - → We will not implement the reduction on a partially updated DMAF, even if 50 of the 51 states have been delivered by January 7, 2000.
  - → Likewise, if there is a reissue of a MAF Extract after January 7, 2000, then we will not redo the reduction. We will stop all work on using an updated DMAF for the reduction, and the December results will be "official".

- On the surface, this contingency plan sounds fairly safe and straightforward to implement. However, the DMAF counts are critical input to the small block subsampling and it is important to have the January DMAF counts for this operation. It is risky to expect that the January results will be verified in time, so the contingency plan must address getting the January DMAF numbers for small block cluster subsampling under the scenario that the December results become official.
  - → The accompanying flowchart shows the general flow of operations.
  - → Both the December and January DMAF counts will be placed on the Sample Design File (SDF).
  - → Regardless of which results are official, the small block cluster subsampling operation will use the January DMAF counts which are on a specified location on the SDF.

A.C.E. Cluster Reduction: Two-Stage Processing Flow

December 14, 1999



UNITED STATES DEPARTMENT OF COMMERCE
Bureau of the Census
Washington, DC 20233-0001

MASTER FILE

February 1, 2000

DSSD CENSUS 2000 PROCEDURES AND OPERATIONS MEMORANDUM SERIES R- 24

MEMORANDUM FOR Maureen Lynch
Assistant Division Chief, Computer Match Processing
Decennial Statistical Studies Division

From: Donna Kostanich DK
Assistant Division Chief, Sampling and Estimation
Decennial Statistical Studies Division

Prepared by: Matt Salganik MS
Decennial Statistical Studies Division

Subject: Accuracy and Coverage Evaluation Survey: Small Block Cluster Subsampling

I. Introduction

This memorandum provides instructions for the small block cluster subsampling operation. Before this operation the Accuracy and Coverage Evaluation (A.C.E.) reduction sample will contain 5,000 small clusters in the United States and 96 small clusters in Puerto Rico. Small clusters are expected to have between zero and two housing units based on an early census address list. Conducting interviewing and follow-up operations in clusters of this size is not as cost-effective as in larger clusters. Therefore, to allocate A.C.E. resources more efficiently, we will only include a subsample of these small clusters in the A.C.E. interviewing sample. The same subsampling procedure will be used for clusters in the United States and Puerto Rico.

This subsampling operation will reduce the sample of small clusters while at the same time attempting to balance among three goals. First, we would like to prevent any small clusters from having weights that are extremely high compared to other clusters in the sample. Second, we would like to have lower weights on clusters where the number of housing units is different than we expected. These first two goals attempt to reduce the contribution of small clusters to the variance of the dual system estimates. The third goal is to ensure that the Field Division can efficiently manage the resulting workloads.

To achieve these goals we will use differential subsampling where the subsampling rates are based on the number of keyed and valid housing units from the A.C.E. Independent Listing[1] (IL) and the number of housing units on the Decennial Master Address File (DMAF). This DMAF

[1] This IL housing unit count includes units with the status 'future new construction.'

housing unit count will be based on the January 2000 update. All American Indian County[2] (AIC), American Indian Reservation (AIR), and List/Enumerate clusters will be retained to avoid increasing the weights on these clusters.

The exact subsampling rates have not yet been determined. See Attachment A for the approximate take-everys. The exact rates will be determined after the keyed and valid IL counts and January DMAF are available. Once the rates have been determined, they will be keyed into a parameter file which will be provided to you in late January of 2000, a few days after we are provided with the keyed and valid housing units counts.

Small block cluster subsampling is part of the larger process of selecting the sample clusters for the A.C.E. This process begins with the listing sample selection which yields a sample of approximately 2 million housing units. An independent listing operation is done to create an address list. Next, the listing sample, which was based on the design of the Integrated Coverage Measurement Survey, will be subsampled to yield the A.C.E. reduced sample, which will be based on the A.C.E. design. Small block cluster subsampling will then occur resulting in the A.C.E. sample clusters. These are the clusters that will be in the A.C.E. interview sample. However, there is one more sampling process before we arrive at our final 300,000 housing unit sample – large block cluster subsampling. In this process some housing units will be removed from sample in large block clusters (those with 80+ housing units). The remaining housing units after large block cluster subsampling make up the A.C.E. interview sample.

This specification should be used to flowchart the process, to generate further discussion on requirements, to identify and finalize record layouts of input and output files, and to write computer software to implement the methodology. During and after a testing phase, it is likely that changes to the specification will be necessary.

Any comments or questions should be directed to Matt Salganik (301-457-3636) or Debbie Fenstermaker (301-457-4195).

II. Assumptions

A. The A.C.E. block cluster reduction has been completed.

B. All independent listing counts used will be 'keyed and valid' counts.

C. For purposes of small block cluster subsampling, the independent list housing unit counts include units coded as 'future new construction'.

---

[2]American Indian Country includes Tribal Jurisdiction Statistical Areas, Tribal Designated Statistical Areas, Alaska Native Village Statistical Areas, and American Indian Reservations and associated trustlands. Throughout this specification the term American Indian Country will exclude American Indian Reservations and associated trustlands.

D. The definition of American Indian Country is as of spring 1999 when block clustering was completed. This information could change before all census operations are completed.

E. Small block cluster subsampling take-everys have been set so that the expected number of clusters from each stratum will be an integer.

III. Process

In this process the block clusters will be put into eleven different small block cluster subsampling strata using information from the Sample Design File and the independent listing results. The A.C.E. clusters from the medium, large, and AIR sampling strata are not part of the small block cluster subsampling process, and therefore will be retained in this operation. However, we still assign these clusters to small block cluster subsampling strata and pick up their independent list housing unit (HU) counts. This information is used in large block cluster subsampling. All small clusters that are AIR, AIC, and List/Enumerate as well as those with 10 or more housing units on either the DMAF or IL will be retained in this process.

All the following steps should only be completed for clusters that are in the A.C.E. reduced sample (Current Sample Indicator = 1 on the Sample Design File).

A. Assigning Clusters to a Small Block Cluster Subsampling Stratum

1. Using the results of the independent list keying operation, obtain the number of IL HUs for each block cluster and write this value to the variable NHUIL on the Sample Design File. See Attachment B for a layout of the Sample Design File. Include HUs with all eight of the Unit Status Codes[3] in this IL count.

2. For each block cluster create a new variable, LARGERHU, which is equal to the larger of the DMAF HU count (from the Sample Design File) and IL HU count. For some block clusters such as List/Enumerate, there will be no DMAF count. In these cases assign the IL HU count to the LARGERHU variable. Write this variable to the Sample Design File.

3. Using the size category (SIZECAT), the American Indian Country Indicator (AICIND), the larger of the DMAF or IL count (LARGERHU)

[3]The Unit Status codes are:

1) Occupied or vacant and intended for occupancy
2) Under construction
3) Future construction
4) Unfit for habitation
5) Boarded up
6) Storage of household goods
7) Vacant mobile home site
8) Other

and the List/Enumerate flag (LEIND) from the Sample Design File, assign each cluster a small block cluster subsampling stratum code based on the following table. Write these values to the variable SBCSS on the Sample Design File.

Table 1. Small Block Cluster Subsampling Strata Assignment Rules

IF				THEN
Original Cluster Size Category (SIZECAT)	Larger of DMAF/IL HU count (LARGERHU)	American Indian Country Indicator[4] (AICIND)	List/ Enumerate Indicator (LEIND)	Sub-Sampling Stratum Code (SBCSS)
Small (0-2)	0-2	0 (non AIR/AIC)	0 (not L/E)	01
			1 (L/E)	05
		1 (AIR)	0 or 1	07
		2 (AIC)	0 or 1	08
Small (0-2)	3-5	0 (non AIR/AIC)	0 (not L/E)	02
			1 (L/E)	06
		1 (AIR)	0 or 1	07
		2 (AIC)	0 or 1	09
Small (0-2)	6-9	0 (non AIR/AIC)	0 (not L/E)	03
			1 (L/E)	06
		1 (AIR)	0 or 1	07
		2 (AIC)	0 or 1	09
Small (0-2)	10+	0, 1, or 2	0 or 1	04
Medium (3-79)	all	0, 1, or 2	0 or 1	10
Large (80+)	all	0, 1, or 2	0 or 1	11

[4]For the American Indian Country Indicator :
0 = Not American Indian Country
1 = American Indian Reservation or Trustland
2 = Tribal Jurisdiction statistical area/Alaska Native Village statistical area/tribal designated statistical area

Stratum 01 – Non L/E Small Block Clusters where 0 ≤ LARGERHU ≤ 2 not on AIR or AIC
Stratum 02 – Non L/E Small Block Clusters where 3 ≤ LARGERHU ≤ 5 not on AIR or AIC
Stratum 03 – Non L/E Small Block Clusters where 6 ≤ LARGERHU ≤ 9 not on AIR or AIC
Stratum 04 – Small Block Clusters where 10 ≤ LARGERHU
Stratum 05 – L/E Small Block Clusters where 0 ≤ LARGERHU ≤ 2 not on AIR or AIC
Stratum 06 – L/E Small Block Clusters where 3 ≤ LARGERHU ≤ 9 not on AIR or AIC
Stratum 07 – American Indian Reservation where 0 ≤ LARGERHU ≤ 9
Stratum 08 – American Indian Country where 0 ≤ LARGERHU ≤ 2
Stratum 09 – American Indian Country where 3 ≤ LARGERHU ≤ 9
Stratum 10 – Medium Block Clusters
Stratum 11 – Large Block Clusters

B. Sending Clusters to Housing Unit Matching

Some of the strata will not be subsampled. Since it is important to start housing unit matching as soon as possible, clusters in these strata should be sent to housing unit matching in a timely manner to prevent any delay in operations.

1. Create a new variable SB on the Sample Design File to indicate whether a cluster has been retained during the small block cluster subsampling operation. Since all the clusters in small block cluster subsampling strata 04, 07, 10, and 11 will be retained in sample, create and set the random start variable (RSSB) for these cluster to 1.000000, the initial and final take-every (ITESB and FTESB) to 1.000000, and set SB equal to one. Then send them to housing unit matching.[5]

2. The small clusters in small block cluster strata 01, 02, 03, 05, 06, 08, and 09 will be subsampled to determine which clusters will be sent to HU matching.

C. Subsample the Small Block Clusters in Strata 01, 02, 03, 05, 06, 08, and 09

For the strata that will be subsampled, the small block cluster subsampling will be done separately for each small block cluster subsampling stratum within each state. The subsampling take-everys will be set so that the expected number of clusters from each small block cluster subsampling stratum is an integer. Subsampling for a specific state should not begin until every small cluster in that state has been assigned to a small block cluster subsampling stratum.

---

[5]It turns out that all clusters in strata 05, 06, 08, and 09 will also be retained in the sample. However, when this specification and the computer programs to implement the operation were being created that was not yet known. We ensure that these clusters are selected by setting their take-everys to one in the small block cluster subsampling parameter file.

1. Sort the block clusters within each small block cluster subsampling stratum (SBCSS) by estimated cluster urbanization (ECLUSURB), county (COUNTY), and A.C.E. cluster number including check digit (CLUST). This sort will help to insure that our sample is representative across these geographic levels.

2. Within each small block cluster subsampling stratum, create an index by numbering the block clusters from 1 to N where N is the number of block clusters in the subsampling stratum.

3. For each small block cluster subsampling stratum, get the initial take-every (ITESB) from the Small Block Subsampling Parameter File that the Sample Design Team has provided. Write this value to the Sample Design File. See Attachment C for a layout.

4. If the number of clusters, N, in a small block cluster subsampling stratum does not equal zero and is less than the initial take-every for that stratum (ITESB) then set the final stratum take-every (FTESB) to the number of clusters in the stratum. Otherwise, if the number of clusters in the small block cluster subsampling stratum is zero or greater than the ITESB set the final stratum take-every (FTESB) equal to the initial stratum take-every (ITESB). This is done to insure that we will select at least one cluster from each small block cluster subsampling stratum.[6]

5. Generate a sequence of numbers $L_1, L_2, ..., L_n$ as follows:

   a. For each subsampling stratum, generate a random number (RN) between 0 and 1 ($0 < RN \leq 1$) with 10 decimal places.

   b. Calculate a random start, RSSB, which equals RN×FTESB. Round this number to six decimal places and write it to the Sample Design File record for each cluster in the subsampling stratum.

   c. Let L_1 = RSSB.

   d. Calculate $L_J = L_{J-1}$ + FTESB for J = 2, 3, ..., n, where n is the largest integer such that [RSSB + (n - 1)×FTESB] $\leq$ N.

---

[6]This step in the specification is no longer necessary because before the take-everys are included in the small block cluster subsampling parameter file they will be computed to ensure that they yield integer expected sample sizes for each subsampling stratum. However, when the specification and computer programs were being written it was not yet known that the take-everys would be computed in this way.

e. Round each L_J up to the nearest integer (an integer rounds to itself).

6. For each cluster in the subsampling stratum with an index equal to the rounded values of L_J, J = 1, 2, ..., n, assign SB = 1. These block clusters are in the sample. Send them on to housing unit matching.

7. For each cluster in the subsampling stratum with an index not equal to the rounded values of L_J, J = 1, 2, ..., n, assign SB = 0. These block clusters are not in the sample. For these clusters set the Current Sample Indicator on the Sample Design File to 0.

8. For each subsampling stratum calculate a check value C such that:

$$C = | (N / FTESB) - n |$$

N = Number of clusters in the subsampling stratum
n = Number of clusters selected from the subsampling stratum
FTESB = Final small block cluster subsampling stratum take-every

If the sampling procedure was performed correctly, then C will be less than one. If C is greater than or equal to one, then contact the author so that operations can be reviewed.[7]

D. Calculate Cluster Weights

For all clusters in the A.C.E. sample after small block cluster subsampling (small, medium, large, and AIR), compute the variable WEIGHTC which is equal to the unbiased weight of each cluster. Calculate this value for each cluster by multiplying the take-everys from the initial block cluster sampling, the take-every from the A.C.E. reduction and the final take-every from small block cluster subsampling. Round to six decimal places and write to the Sample Design File. For all clusters not in the A.C.E. sample after small block cluster subsampling leave this value blank.

WEIGHTC = TE1 × TE2 × TEAR × FTESB

[7]Before the take-everys were written to the small block clusters subsampling parameter file they were computed to yield integer expected sample sizes from all the subsampling strata. Because of this, C will always be equal to 0. This was not written into the specification because at the time of its writing we did not know the take-everys would be computeded in this way.

E. Produce Verification Output

1. Provide the sampling staff with access to the Cluster Status File, so that with the Sample Design File we may replicate the sampling operation.

2. Create the Block Cluster Subsampling Verification File. This file provides summary information for the different strata in each state and will be used by the sampling staff for verification. See Attachment D for a layout. Several calculations are required for this file.

   a. To calculate the average LARGERHU for all clusters in the stratum add up the LARGERHU values in the stratum and divide by the number of clusters in the stratum.

   b. To calculate the average LARGERHU of clusters selected from a stratum add up the LARGERHU values of the clusters selected from the stratum and divide by the number of clusters selected from the stratum.

   c. For small block cluster subsampling strata one through nine, calculate the weight of a cluster selected from a stratum by multiplying TE1 × TE2 × TEAR × FTESB. For strata one through nine the weights of each cluster within a stratum within a state should be equal.

3. Create the Independent List Housing Unit Information File. Please provide us this file as soon as possible so that we may set our final take-everys. This file is required in the setting of the take-everys because we want to insure that the expected number of clusters from each small block cluster subsampling stratum is an integer. This file will also be used by the sampling staff for evaluation purposes. See Attachment E for a layout.

IV. Input

The following files are inputs to this operation.

A. Cluster Status File

This file contains one record for each block cluster selected in the A.C.E. listing sample. The original source of this file is the Sample Design File. It is updated with information from other processing stages. For small block cluster subsampling, this file is used to obtain the keyed and valid counts of Independent Listing housing units by type of unit status. For information about this file contact Courtney Ford in the Processing Support & A.C.E. Systems Staff at 301-457-4121.

B. Sample Design File, Version 2

This file contains information about the entire sampling history of each block cluster. See Attachment B for a layout. Version 2 reflects listing sample selection and A.C.E. cluster reduction. There are a total of 29,717 cluster records on the file. Only clusters with CSI= 1 are in the reduced A.C.E. sample. Note that once a cluster drops out of sample, the fields from the remaining operations will be left blank.

C. Small Block Cluster Subsampling Parameter File

This file, which contains one record for each state, records the initial take-every for each of the eleven small block cluster subsampling strata. These initial take-everys will be set so that the expected sample size from each small block cluster subsampling strata will be an integer. See Attachment C for a layout. This file will be created in late January by the Sample Design Team.

V. Output

A. Sample Design File, Version 3

This file contains information about the entire sampling history of each block cluster. See Attachment B for a layout. It will be updated after the small block cluster subsampling process.

B. Small Block Cluster Subsampling Verification File

This file will be used to assist the sample design staff in verification procedures. See Attachment D for a layout.

C. Independent Listing Housing Unit Information File

This file will be used by the sample design staff to evaluate the stratification of the small block clusters. It will contain one record for each cluster (including medium, large, and AIR clusters) still in sample before Small Block Cluster Subsampling. See Attachment E for a layout.

cc: DSSD Census 2000 Procedures and Operations Memorandum Series Distribution List
A.C.E. Team Leaders
Statistical Design Team Leaders
Sample Design Team
Estimation Team
Variance Estimation Team

Approximate Small Block Cluster Subsampling Take-Everys

Listed below are the approximate small block cluster subsampling take-everys. These take-everys have not been computed to insure that the expected number of clusters from each stratum is an integer. This computation takes place before the take-everys are written to the small block cluster subsampling parameter file, and it is not part of this specification. The final take-everys, as well as the exact methodology for their calculation, are forthcoming in a later document. Because the final take-everys will be different than the ones listed below, this attachment is included merely as a guide.

FIPS CODE	State Name	Stratum 1	Stratum 2	Stratum 3	Stratum 4-9
01	Alabama	10.00	4.00	2.22	1.00
02	Alaska	10.00	4.00	2.22	1.00
04	Arizona	2.80	1.12	1.00	1.00
05	Arkansas	8.97	3.59	1.99	1.00
06	California	2.57	1.03	1.00	1.00
08	Colorado	4.11	1.64	1.00	1.00
09	Connecticut	8.71	3.48	1.93	1.00
10	Delaware	10.00	4.00	2.22	1.00
11	District of Columbia	10.00	4.00	2.22	1.00
12	Florida	4.66	1.86	1.04	1.00
13	Georgia	10.00	4.00	2.22	1.00
15	Hawaii	3.92	1.57	1.00	1.00
16	Idaho	2.09	1.00	1.00	1.00
17	Illinois	10.00	4.00	2.22	1.00
18	Indiana	10.00	4.00	2.22	1.00
19	Iowa	10.00	4.00	2.22	1.00
20	Kansas	10.00	4.00	2.22	1.00
21	Kentucky	10.00	4.00	2.22	1.00
22	Louisiana	3.59	1.43	1.00	1.00
23	Maine	6.24	2.50	1.38	1.00
24	Maryland	9.48	3.79	2.11	1.00
25	Massachusetts	8.47	3.39	1.88	1.00
26	Michigan	9.22	3.69	2.05	1.00
27	Minnesota	10.00	4.00	2.22	1.00
28	Mississippi	6.80	2.72	1.51	1.00
29	Missouri	10.00	4.00	2.22	1.00
30	Montana	2.68	1.07	1.00	1.00
31	Nebraska	10.00	4.00	2.22	1.00
32	Nevada	1.90	1.00	1.00	1.00
33	New Hampshire	10.00	4.00	2.22	1.00
34	New Jersey	5.51	2.20	1.22	1.00
35	New Mexico	2.16	1.00	1.00	1.00

36	New York	9.94	3.98	2.21	1.00
37	North Carolina	10.00	4.00	2.22	1.00
38	North Dakota	8.10	3.24	1.80	1.00
39	Ohio	10.00	4.00	2.22	1.00
40	Oklahoma	9.70	3.88	2.16	1.00
41	Oregon	2.11	1.00	1.00	1.00
42	Pennsylvania	10.00	4.00	2.22	1.00
44	Rhode Island	10.00	4.00	2.22	1.00
45	South Carolina	10.00	4.00	2.22	1.00
46	South Dakota	8.63	3.45	1.92	1.00
47	Tennessee	10.00	4.00	2.22	1.00
48	Texas	3.20	1.28	1.00	1.00
49	Utah	2.17	1.00	1.00	1.00
50	Vermont	10.00	4.00	2.22	1.00
51	Virginia	9.03	3.61	2.01	1.00
53	Washington	3.00	1.20	1.00	1.00
54	West Virginia	7.05	2.82	1.57	1.00
55	Wisconsin	10.00	4.00	2.22	1.00
56	Wyoming	1.94	1.00	1.00	1.00
72	Puerto Rico	3.31	1.32	1.00	1.00

Sample Design File Layout

The Sample Design File contains one record per block cluster selected during the initial block cluster sampling. If the block cluster is subsampled out of sample during the second step of sampling, the A.C.E. reduction or during small block subsampling, the remaining variables will be left blank. The initial version of the file, which will be created following the initial block cluster selection, is called SDF.US1. For each subsequent update to the file, the version number will increase by one (i.e. SDF.US2, SDF.US3). The layout for the Sample Design File is as follows:

Variable Description	Name	Places	Source
Census Region	REGION	1	UN
Census Division	DIV	2	UN
State code	STATE	3-4	UN
County code	COUNTY	5-7	UN
Local census office	LCO	8-11	CS
Interim Tract (Pseudo Tract)	ITRACT	12-17	BC
Current Sample Indicator	CSI	19	UO
A.C.E. block cluster number	CLUST	21-25	CS
Check Digit	DIGIT	26	CS
Geography block cluster number	GCLUST	28-32	BC
List/Enumerate Indicator (1= L/E, 0 = Non-L/E)	LEIND	33	BC
Type of Enumeration Area Recode	TEACR	34	CS
Type of Enumeration Area group	TEAG	36	BC
Number of HUs used for sample design	NHU	37-41	BC
Number of MAF HUs	NHUM	43-47	BC
Number of 1990 HUs	NHU90	49-53	BC
Sampling Stratum 1 = Small 2 = Medium 3 = Large 4 = American Indian Reservation	SS	55	UN
American Indian Country Indicator 0 = No American Indian Country 1 = American Indian Reservation/trust land 2 = Tribal Jurisdiction Statistical Area/ Tribal Designated Statistical Area/ Alaska Native Village Statistical Area	AICIND	56	BC
Demographic/Tenure Group code	DTCODE	57-58	UN
Demographic/Tenure Group label	DTLABEL	59-60	UN
Estimated Urbanicity of block cluster 1 = Urban Area with population ≥250,000 2 = Other Urban Area 3 = Non-Urban Area	ECLUSURB	62	UN

Variable Description	Name	Places	Source
Size Category	SIZCAT	63	UN
1=Small (0-2 hus)			
2=Medium (3-79 hus)			
3=Large (80+ hus)			
Additional space		64-91	
First step index number	INDEX1	92-99	CS
Listing sample selection indicator	BC1	101	CS
1 = Selected			
Random start for listing sample selection	RS1	103-113	UN
Take-every for listing sample selection	TE1	115-125	UN
Second step listing sample selection indicator	BC2	127	CS
0 = Not Selected, 1 = Selected			
Random start for second step listing sample selection	RS2	129-139	CS
Take-every for second step listing sample selection	TE2	141-151	CS
Unbiased weight after listing sample selection	WEIGHTBC	153-164	CS
Additional space		165-175	
Preliminary Number of HUs on the Independent List	NHUILP	176-180	AR
Number of Housing Units On January 2000 DMAF	NHUDMAF	182-186	AR
Demographic code	DEMCODE	188-188	AR
1 = Minority			
2 = Non-minority			
3 = Puerto Rico			
Consistency Code	CONCODE	189-189	AR
1 = Low Inconsistent (IL significantly smaller than DMAF)			
2 = Consistent			
3 = High Inconsistent (IL significantly larger than DMAF)			
A.C.E. reduction stratum	ARS	190-191	AR
A.C.E. reduction indicator	ACERED	193-193	AR
0 = Not Selected, 1 = Selected			
Random start for A.C.E. reduction	RSAR	195-205	AR
Take-every for A.C.E. reduction	TEAR	207-217	AR
Unbiased weight after A.C.E. reduction	WEIGHTAR	219-230	AR
Collapsing flag	COLFLAG	232-232	AR
A.C.E. Reduction index number	INDEXR	234-241	AR
Number of Housing Units on the December 1999 DMAF (Initial)	NHUDMAFI	243-247	AR
Additional space		248-300	
Number of HUs on the Independent List	NHUIL	301-305	SB
Small Block Cluster Subsampling Stratum	SBCSS	306-307	SB
Small Block Subsampling Indicator	SB	308	SB
0 = Not Selected, 1 = Selected			
Random Start for Small Block subsampling	RSSB	310-320	SB
Initial take-every for Small Block subsampling	ITESB	322-332	SB
Unbiased weight for ACE cluster	WEIGHTC	334-345	SB
Larger of the DMAF and IL HU count	LARGERHU	347-351	SB
Final take-every for Small Block subsampling	FTESB	352-362	SB
Additional space		363-370	SB

<u>Source Codes</u>

AR: ACE Reduction
BC: Block Clustering
CS: Block Cluster Sampling
SB: Small Block Subsampling
UN: Universe File Creation
UO: Updated for each operation

Small Block Cluster Subsampling Parameter File Layout

This file, which will be created by the sample design staff, will provide the take-every for the different strata in each state. There will be one record for each state, the District of Columbia, and Puerto Rico. It will be called SBCSPF.DAT.

Variable Description	Name	Places
State (FIPS Code)	STATE	1-2
Take-every for stratum 01*	TE_1	4-14
Take-every for stratum 02*	TE_2	16-26
Take-every for stratum 03*	TE_3	28-38
Take-every for stratum 04*	TE_4	40-50
Take-every for stratum 05*	TE_5	52-62
Take-every for stratum 06*	TE_6	64-74
Take-every for stratum 07*	TE_7	76-86
Take-every for stratum 08*	TE_8	88-98
Take-every for stratum 09*	TE_9	100-110
Take-every for stratum 10*	TE_10	112-122
Take-every for stratum 11*	TE-11	124-134

*Note: Take-everys will be rounded to six decimal places and may be non-integer values.

Small Block Cluster Subsampling Verification File Layout

This file will be created during processing to assist the sample design staff in verification. One file will be created for the entire nation (including the District of Columbia and Puerto Rico). This file will be called SBCSVF.DAT.

Variable Description	Name	Places
State (FIPS Code)	STATE	1-2
Initial take-every for stratum 01	ITE_1	4-14
Number of clusters in stratum 01	CLUS_1	15-17
Final take-every for stratum 01	FTE_1	18-28
Number of clusters selected from stratum 01	SCLUS_1	29-31
Number of IL HUs in stratum 01	ILHU_1	32-36
Number of IL HUs in clusters selected from stratum 01	SILHU_1	37-41
Average LARGERHU of clusters in stratum 01 *Rounded to three decimal places*	ALHU_1	42-47
Average LARGERHU of clusters selected from stratum 01 *Rounded to three decimal places*	SALHU_1	48-53
Weight of clusters selected from stratum 01	WEIGHT_1	54-65
Random number used to sample stratum 01	RN_1	66-76
Initial take-every for stratum 02	ITE_2	88-98
Number of clusters in stratum 02	CLUS_2	99-101
Final take-every for stratum 02	FTE_2	102-112
Number of clusters selected from stratum 02	SCLUS_2	113-115
Number of IL HUs in stratum 02	ILHU_2	116-120
Number of IL HUs in clusters selected from stratum 02	SILHU_2	121-125
Average LARGERHU of clusters in stratum 02 *Rounded to three decimal places*	ALHU_2	126-131
Average LARGERHU of clusters selected from stratum 02 *Rounded to three decimal places*	SALHU_2	132-137
Weight of clusters selected from stratum 02	WEIGHT_2	138-149
Random number used to sample stratum 02	RN_2	150-160
Initial take-every for stratum 03	ITE_3	182-192
Number of clusters in stratum 03	CLUS_3	193-195
Final take-every for stratum 03	FTE_3	196-206
Number of clusters selected from stratum 03	SCLUS_3	207-209
Number of IL HUs in stratum 03	ILHU_3	210-214
Number of IL HUs in clusters selected from stratum 03	SILHU_3	215-219
Average LARGERHU of clusters in stratum 03 *Rounded to three decimal places*	ALHU_3	220-225
Average LARGERHU of clusters selected from stratum 03 *Rounded to three decimal places*	SALHU_3	226-231
Weight of clusters selected from stratum 03	WEIGHT_3	232-243
Random number used to sample stratum 03	RN_3	244-254

Initial take-every for stratum 04	ITE_4	286-296
Number of clusters in stratum 04	CLUS_4	297-299
Final take-every for stratum 04	FTE_4	300-310
Number of clusters selected from stratum 04	SCLUS_4	311-313
Number of IL HUs in stratum 04	ILHU_4	314-318
Number of IL HUs in clusters selected from stratum 04	SILHU_4	319-323
Average LARGERHU of clusters in stratum 04 *Rounded to three decimal places*	ALHU_4	324-329
Average LARGERHU of clusters selected from stratum 04 *Rounded to three decimal places*	SALHU_4	330-335
Weight of clusters selected from stratum 04	WEIGHT_4	336-347
Initial take-every for stratum 05	ITE_5	390-400
Number of clusters in stratum 05	CLUS_5	401-403
Final take-every for stratum 05	FTE_5	404-414
Number of clusters selected from stratum 05	SCLUS_5	415-417
Number of IL HUs in stratum 05	ILHU_5	418-422
Number of IL HUs in clusters selected from stratum 05	SILHU_5	423-427
Average LARGERHU of clusters in stratum 05 *Rounded to three decimal places*	ALHU_5	428-433
Average LARGERHU of clusters selected from stratum 05 *Rounded to three decimal places*	SALHU_5	434-439
Weight of clusters selected from stratum 05	WEIGHT_5	440-451
Random number used to sample stratum 05	RN_5	452-462
Initial take-every for stratum 06	ITE_6	494-504
Number of clusters in stratum 06	CLUS_6	505-507
Final take-every for stratum 06	FTE_6	508-518
Number of clusters selected from stratum 06	SCLUS_6	519-521
Number of IL HUs in stratum 06	ILHU_6	522-526
Number of IL HUs in clusters selected from stratum 06	SILHU_6	527-531
Average LARGERHU of clusters in stratum 06 *Rounded to three decimal places*	ALHU_6	532-537
Average LARGERHU of clusters selected from stratum 06 *Rounded to three decimal places*	SALHU_6	538-543
Weight of clusters selected from stratum 06	WEIGHT_6	544-555
Random number used to sample stratum 06	RN_6	556-566

Initial take-every for stratum 07	ITE_7	598-608
Number of clusters in stratum 07	CLUS_7	609-611
Final take-every for stratum 07	FTE_7	612-622
Number of clusters selected from stratum 07	SCLUS_7	623-625
Number of IL HUs in stratum 07	ILHU_7	626-630
Number of IL HUs in clusters selected from stratum 07	SILHU_7	631-635
Average LARGERHU of clusters in stratum 07 *Rounded to three decimal places*	ALHU_7	636-641
Average LARGERHU of clusters selected from stratum 07 *Rounded to three decimal places*	SALHU_7	642-647
Weight of clusters selected from stratum 07	WEIGHT_7	648-659
Initial take-every for stratum 08	ITE_8	702-712
Number of clusters in stratum 08	CLUS_8	713-715
Final take-every for stratum 08	FTE_8	716-726
Number of clusters selected from stratum 08	SCLUS_8	727-729
Number of IL HUs in stratum 08	ILHU_8	730-734
Number of IL HUs in clusters selected from stratum 08	SILHU_8	735-739
Average LARGERHU of clusters in stratum 08 *Rounded to three decimal places*	ALHU_8	740-745
Average LARGERHU of clusters selected from stratum 08 *Rounded to three decimal places*	SALHU_8	746-751
Weight of clusters selected from stratum 08	WEIGHT_8	752-763
Random number used to sample stratum 08	RN_8	764-774
Initial take-every for stratum 09	ITE_9	806-816
Number of clusters in stratum 09	CLUS_9	817-819
Final take-every for stratum 09	FTE_9	820-830
Number of clusters selected from stratum 09	SCLUS_9	831-833
Number of IL HUs in stratum 09	ILHU_9	834-838
Number of IL HUs in clusters selected from stratum 09	SILHU_9	839-843
Average LARGERHU of clusters in stratum 09 *Rounded to three decimal places*	ALHU_9	844-849
Average LARGERHU of clusters selected from stratum 09 *Rounded to three decimal places*	SALHU_9	850-855
Weight of clusters selected from stratum 09	WEIGHT_9	856-867
Random number used to sample stratum 09	RN_9	868-878

Initial take-every for stratum 10	ITE_10	910-920
Number of clusters in stratum 10	CLUS_10	921-923
Final take-every for stratum 10	FTE_10	924-934
Number of clusters selected from stratum 10	SCLUS_10	935-937
Number of IL HUs in stratum 10	ILHU_10	938-942
Number of IL HUs in clusters selected from stratum 10	SILHU_10	943-947
Average LARGERHU of clusters in stratum 10 *Rounded to three decimal places*	ALHU_10	948-953
Average LARGERHU of clusters selected from stratum 10 *Rounded to three decimal places*	SALHU_10	954-959
Initial take-every for stratum 11	ITE_11	1014-1024
Number of clusters in stratum 11	CLUS_11	1025-1027
Final take-every for stratum 11	FTE_11	1028-1038
Number of clusters selected from stratum 11	SCLUS_11	1039-1041
Number of IL HUs in stratum 11	ILHU_11	1042-1046
Number of IL HUs in clusters selected from stratum 11	SILHU_11	1047-1051
Average LARGERHU of clusters in stratum 11 *Rounded to three decimal places*	ALHU_11	1052-1058
Average LARGERHU of clusters selected from stratum 11 *Rounded to three decimal places*	SALHU_11	1059-1065

Independent List Housing Unit Information File Layout

This cluster level file will be created during processing to assist the sample design staff in evaluation of the subsampling. It may also be used to help verify the large block cluster subsampling parameters. One file will be created for the nation (including the District of Columbia and Puerto Rico). The file will include one record for each cluster (including medium, large, and AIR clusters) in sample before Small Block Cluster Subsampling. The counts on this file will be keyed and valid IL counts. The file will be called ILHUIF.DAT.

Variable Description	Name	Places
State (FIPS Code)	STATE	1-2
County	COUNTY	3-5
A.C.E. Cluster Number	CLUST	6-10
Check Digit	DIGIT	11-11
Total number of IL HUs	ILHU	12-16
Number of HUs where USTAT = 1 *(Occupied or vacant and intended for occupancy)*	USTAT_1	17-21
Number of HUs where USTAT = 2 *(Under construction)*	USTAT_2	22-26
Number of HUs where USTAT = 3 *(Future Construction)*	USTAT_3	27-31
Number of HUs where USTAT = 4 *(Unfit for Habitation)*	USTAT_4	32-36
Number of HUs where USTAT = 5 *(Boarded Up)*	USTAT_5	37-41
Number of HUs where USTAT = 6 *(Storage of household goods)*	USTAT_6	42-46
Number of HUs where USTAT = 7 *(Vacant mobile home site)*	USTAT_7	47-51
Number of HUs where USTAT = 8 *(Other)*	USTAT_8	52-56

UNITED STATES DEPARTMENT OF COMMERCE
Bureau of the Census
Washington, DC 20233-0001

March 8, 2000

MASTER FILE

DSSD CENSUS 2000 PROCEDURES AND OPERATIONS MEMORANDUM SERIES R-26

MEMORANDUM FOR Maureen Lynch
Assistant Division Chief, Coverage Measurement Processing
Decennial Statistical Studies Division

From: Donna Kostanich
Assistant Division Chief, Sampling and Estimation
Decennial Statistical Studies Division

Prepared by: Ryan Cromar RC
Sample Design Team
Decennial Statistical Studies Division

Subject: Accuracy and Coverage Evaluation: Large Block Cluster Subsampling Parameter File Specifications

I. INTRODUCTION

This memorandum provides specifications for the large block cluster subsampling parameter file for the Census 2000 Accuracy and Coverage Evaluation (A.C.E.) survey. As the final stage of the A.C.E. design, large block cluster subsampling involves selecting a portion of a block cluster that has 80 or more A.C.E. housing units to be in the A.C.E. interview sample. This will be accomplished by forming segments of adjacent housing units within the block cluster and selecting a subsample of segments. The objective of large block cluster subsampling is to meet the target A.C.E. interviewing sample sizes using the most up-to-date A.C.E. housing unit counts available at the time of subsampling. The creation of the parameter file is the first step of large block cluster subsampling. The parameter file is an input for the large block cluster subsampling process documented in reference 1. The large block cluster subsampling parameter file specification is similar to the specifications prepared for the 1998 Census 2000 Dress Rehearsal and documented in reference 2.

Earlier stages of the A.C.E. sample design include the selection of A.C.E. block clusters for the listing sample (see reference 3), the A.C.E. reduction (see reference 4), and the subsampling of small block clusters (see reference 5). After the listing sample selection, the independent list is created and the results keyed and verified. Based on the results of this listing, the A.C.E. sample reduction and small block cluster subsampling are done. Subsequently, the housing unit matching and follow-up are done, and the preliminary enhanced list is created and sent to large block cluster subsampling. The preliminary enhanced list is both the input and output file for large block cluster subsampling. The output preliminary enhanced list is updated with the subsampling results, and is referred to as the subsampled preliminary enhanced list. The enhanced List is created by extracting only the housing units designated for interview following large block cluster subsampling from the subsampled preliminary enhanced list (see reference 6).

This memorandum is organized into the following sections:

- Assumptions
- Definitions
- Input
- Process
- Output
- Verification
- References

These specifications should be used to flowchart the process, to generate further discussion on requirements, to identify and finalize the record layouts of input and output files, and to write computer software to implement the methodology. During and after a testing phase, it is possible that changes to the specifications will be necessary.

If there are any questions or comments, please contact Ryan Cromar (301-457-1636), James Farber (301-457-4282), or Deborah Fenstermaker (301-457-4195) of the Decennial Statistical Studies Division (DSSD).

II. ASSUMPTIONS

The assumptions required to create the large block cluster subsampling parameter file are:

A. The creation of the large block cluster subsampling parameter file is not affected by the results of housing unit matching and follow-up. Thus the creation of this parameter file can be done using independent list housing unit counts at any time after small block cluster subsampling even if matching and follow-up are not completed. It is possible that the housing unit counts determined in this specification will differ due to housing unit follow-up.

B. Block clusters eligible for large block cluster subsampling include those that were selected in the A.C.E. reduction and remain in the sample following small block cluster subsampling. All other block clusters are not eligible for large block cluster subsampling.

C. Large block cluster subsampling is done on a flow basis over a span of several days. Therefore, the large block cluster subsampling parameter file includes fields for daily statistics that will be filled during the subsampling process and will be used to track the day-to-day results of large block cluster subsampling.

D. The large block cluster subsampling parameter file will not be revised to account for relisted block clusters. Relisted block clusters will receive the original subsampling parameters computed from the independent list before relisting. Block clusters which require relisting will not be identified nor will the relisting be done by the time the large block cluster subsampling parameters are calculated.

E. The A.C.E. housing units on the independent list are keyed and valid.

F. The A.C.E. housing units that have a Unit Status of Future Construction are excluded from the take-every calculation. Including such units could cause the A.C.E. interview sample size to be lower than expected, which would increase the variance of the A.C.E. population estimates. Excluding these units is a conservative approach to ensure that target sample sizes are achieved.

G. All decimal numbers are rounded to six digits at the time of creation using the standard rounding procedure except when noted otherwise. Decimal numbers with a seventh decimal place of five or more are rounded up in the sixth decimal place. Those with four or less in the seventh decimal place are rounded down in the sixth decimal place.

H. Note that medium and small block clusters are eligible for large block cluster subsampling since the decision to subsample is based only on the number of A.C.E. housing units in the block cluster.

III. DEFINITIONS

A. American Indian Reservation (AIR) Block Cluster

A block cluster with three or more housing units based on information available at the time block clusters were formed that is at least partially in an AIR. See Sampling Strata.

B. Block Cluster

A geographically contiguous group of Census 2000 collection blocks (see reference 7).

C. Listing Sample

The initial sampling stage of A.C.E. in which block clusters are selected for independent listing (see reference 3).

D. A.C.E. Independent List

List of all housing units in A.C.E. listing sample block clusters. The independent list is created independently of the Decennial Master Address File, the address list used for the census.

E. Keyed and Valid Housing Units

Housing units with any of the following Unit Status codes on the independent list:

1 = Occupied or vacant and intended for occupancy
2 = Under construction
3 = Future construction
4 = Unfit for habitation
5 = Boarded up
6 = Storage of household goods
7 = Vacant mobile home site
8 = Other

All of these units are included because it is possible that the unit status may change between listing and interviewing. Group quarters are not listed in A.C.E. Note that A.C.E. Future construction housing units (Unit Status = 3) are excluded from the take-every calculation as explained in section II above.

F. Large Block Cluster

See Sampling Strata.

G. Medium Block Cluster

See Sampling Strata.

H. A.C.E. Housing Unit

A housing unit on the preliminary enhanced list that is keyed and valid and has one of the following After Follow-up Match Codes: M, MU, UI, or CI.

I. A.C.E. Reduction

The process of reducing the A.C.E. listing sample from the Integrated Coverage Measurement (ICM) sample to the A.C.E. interview sample. In the A.C.E. reduction, the listing sample block clusters are subsampled, and the selected block clusters continue to small block cluster subsampling (see Reference 4).

J. A.C.E. Reduction Strata

A partition (mutually exclusive and exhaustive set) of all block clusters in a state into groups according to certain characteristics. See Attachment A for a list of the A.C.E. Reduction Strata and see reference 4 for more information on how A.C.E. reduction strata are defined.

K. Sampling Strata

A partition of all block clusters within a state into groups according to the number of housing units estimated in each cluster at the time of block clustering (see reference 7). Block Clusters are assigned to sampling strata prior to listing sample selection. The sampling strata are:

1	=	small block clusters with 0 - 2 estimated housing units
2	=	medium non-AIR block clusters with 3 - 79 estimated housing units
3	=	large non-AIR block clusters with ≥ 80 estimated housing units
4	=	medium and large AIR block clusters with ≥ 3 estimated housing units

L. Small Block Cluster

See Sampling Strata.

M. State

The 50 United States plus the District of Columbia and Puerto Rico.

IV. INPUT FILES

The inputs for the subsampling process are the following:

A. Large Block Cluster Subsampling Input File

Description:	This file contains the target housing unit sample size for each A.C.E. reduction stratum within each state.
Level:	A.C.E. reduction stratum
Scope:	One record per A.C.E. reduction stratum within each state
Layout:	See Attachment B

B. Cluster Status File

Description:	This file has one record for each of the 29,695 block clusters selected in the A.C.E. listing sample. It is updated with information from other processing stages. For large block cluster subsampling, this file is used to determine the sampling parameters.
Level:	A.C.E. Block Cluster
Scope:	All block clusters selected for the A.C.E. listing sample

C. A.C.E. Sample Design File (Version 3)

Description:	This file reflects the previous A.C.E. sampling operations: listing sample selection, A.C.E. reduction, and small block cluster subsampling.
Level:	Block Cluster
Scope:	One record for each block cluster in the A.C.E. listing sample
File Layout:	See Attachment C

V. PROCESS

Using results from the independent list, subsampling parameters for each A.C.E. reduction stratum within each state are calculated prior to subsampling any block cluster. The sampling parameters to calculate are the take-every, the target number of segments in a block cluster, and the random start. All block clusters in a common A.C.E. reduction stratum within each state have the same parameters.

A. Determine the take-every by computing the number of keyed and valid housing units on the independent list in all block clusters with 80 or more A.C.E. housing units and dividing it by the target sample size from the block clusters with 80 or more housing units, but excluding the number of future construction housing units from the calculation. Block clusters from the small sampling stratum with fewer than 10 housing units on the independent list are not included when calculating housing unit totals because these housing units are not included in the target sample size.

Use the following steps to determine the take-every for each A.C.E. reduction stratum with each state:

1. Obtain four housing unit counts for each A.C.E. reduction stratum within each state from the cluster status file. Exclude Future Construction housing units, those with Unit Status = 3, from all of these counts.

   a. The number of housing units in all block clusters from the independent list, NILHUT.

   b. The number of housing units in block clusters with 80 or more housing units from the independent list, NILHUL.

   c. The number of housing units in small block clusters (sampling stratum = 1) with fewer than ten housing units from the independent list, NILHUS.

   d. The number of housing units in block clusters with fewer than 80 housing units from the independent list excluding block clusters from step c above, NILHUM.

   e. Calculate Z to check the counts above. This is calculated by subtracting NILHUL, NILHUS and NILHUM from NILHUT.

   $$Z = NILHUT - (NILHUL + NILHUS + NILHUM)$$

   If the counts are correct, Z will be equal to zero. Resolve the cases where Z is not equal to zero.

2. Obtain the target number of sample housing units, T, for the A.C.E. reduction stratum within state from the large block cluster subsampling input file ACE2000_LBINPUT.FIN.

3. Calculate the take-every, TELB:

$$\text{TELB} = \frac{\text{NILHUL}}{\text{T - NILHUM}}$$

- If TELB < 1.012660[1], then set TELB = 1.000000.

- If T - NILHUM = 0, then contact the Sample Design Team of the DSSD.

4. Round the take-every to six decimal places.

B. Calculate the integer number of segments to be formed in a block cluster, NSEG, using the formulas below. These formulas are based on the TELB to ensure at least one segment is selected from each block cluster eligible for subsampling. In addition, create a variable called FORMULA, that codes which formula was used for calculating the number of segments.

- If TELB ≥ 2, then
  NSEG = TELB (If not an integer, round up to the next integer.)
  FORMULA = 1

- If 1 < TELB < 2, then

$$\text{NSEG} = \frac{1}{1 - \frac{1}{\text{TELB}}}$$ (If not an integer, round up to the next integer.)

  FORMULA = 2

- If TELB = 1, then
  NSEG = 1
  FORMULA = 3

[1]This value for TELB was determined to prevent any block clusters from having over 80 segments.

C. Calculate the random start, RS, by generating a random number, RN, between zero and one, rounding it to six decimal places, and multiplying it by the TELB. Round the resulting random start to six decimal places. Calculate a new random start for each A.C.E. reduction stratum within each state.

- $RS = RN \times TELB$, where $0 < RN \leq 1$
- If TELB = 1.000000, then set RS = 1.000000.

D. Starting with the large block cluster subsampling input file as a basis, create a large block cluster sampling parameter file by appending the variables created in this section plus two other variables described in step 1 below. This large block cluster subsampling parameter file will be a daily input to the large block cluster subsampling. Additional variables will be appended during future steps in the large block cluster subsampling process. This file has one record for each A.C.E. reduction stratum within each state.

1. Create two variables, current daily start, DS, and cumulative cluster count, CCC, which are needed for implementing the subsampling over several days. Initialize these variables by setting DS equal to the random start, RS, and assigning a value of zero to CCC for each A.C.E. reduction stratum within each state.

2. Update the following variables on the large block cluster subsampling parameter file using the layout in Attachment D:

   Number of housing units in block clusters with 80 or more housing units on the independent list, NILHUL
   Number of housing units in block clusters with 0-79 housing units on the independent list (except smalls with 0-9), NILHUM
   Number of housing units in all block clusters on the independent list, NILHUT
   Number of housing units in small block clusters with 0-9 housing units on the independent list, NILHUS
   Take-every for the segment subsampling, TELB
   Number of segments in a block cluster, NSEG
   Flag for formula used for calculating NSEG, FORMULA
   Random Number between 0 and 1, RN
   Random Start for the segment subsampling, RS
   Current daily start for the segment subsampling, DS
   Cumulative Cluster Number, CCC

   Round RS, DS and TELB to six decimal places. The other variables are integers.

E. As soon as the parameter file is created, provide it to the Sample Design Team in the DSSD for review and approval prior to large block cluster subsampling.

VI. OUTPUT FILE

The output requested by the Sample Design Team in the DSSD from the Process section is the following:

A. The Large Block Cluster Subsampling Parameter File

Description:	This file contains information needed for selecting the systematic subsample on a flow basis. The file will be produced after the sampling parameters are calculated. The final version will be created when the large block cluster subsampling process is complete.
Level:	A.C.E. Reduction Stratum
Scope:	One record per A.C.E. reduction stratum within each state
File Layout:	See Attachment D

VII. VERIFICATION

The following information should be provided for verification:

- Large Block Cluster Subsampling Parameter File
- Cluster Status File

Provide the sampling parameter file. See section VI above for information about this file. Access to the cluster status file is also required for verification.

VIII. REFERENCES

1 DSSD Census 2000 Procedures and Operations Memorandum Series R-27, "Census 2000 Accuracy and Coverage Evaluation: Large Block Cluster Subsampling Specifications," March 8, 2000.

2 DSSD Census 2000 Dress Rehearsal Memorandum Series A-9, "Census 2000 Dress Rehearsal ICM Sampling: Large Block Subsampling Specification," April 15, 1998.

3 DSSD Census 2000 Procedures and Operations Memorandum Series R-3, "Accuracy and Coverage Evaluation (ACE) Survey: Block Cluster Sample Selection Specification," March 29, 1999.

4 DSSD Census 2000 Procedures and Operations Memorandum Series R-, "Accuracy and Coverage Evaluation Survey: Reduction Specification," January 10, 2000, DRAFT.

5 DSSD Census 2000 Procedures and Operations Memorandum Series R-24, "Accuracy and Coverage Evaluation Survey: Small Block Cluster Subsampling," February 1, 1999.

6 DSSD Census 2000 Procedures and Operations Memorandum Series, Chapter S-HU-08, "Creation of the Census 2000 Accuracy and Coverage Evaluation (A.C.E.) Enhanced List for Person Phase Interviewing," June 21, 1999, DRAFT.

7 DSSD Census 2000 Procedures and Operations Memorandum Series R-8, "Census 2000 Specifications for Block Cluster Formation-Reissue," May 3, 1999.

cc: DSSD Census 2000 Procedures and Operations Memorandum Series Distribution List
A.C.E. Implementation Team Leaders
Statistical Design Team Leaders
Sample Design Team

A.C.E. Reduction Strata

Stratum Code[1]	Stratum Name
01	Minority
02	Non-minority Low Inconsistent
03	Non-minority Consistent
04	Non-minority High Inconsistent
05	Non-minority Inconsistent
06	Non-minority
07	Low Inconsistent
08	Consistent
09	High Inconsistent
10	Inconsistent
11	Minority Inconsistent
12	Minority Consistent
13	Full Collapse
14	Minority Low Inconsistent
15	Minority High Inconsistent
16	Medium Stratum Jumpers
17	American Indian Reservations
18	Puerto Rico
19	Small Stratum Jumpers

[1] Only Strata 01, 02, 03, 04, 16, 17, 18, and 19 were actually used for the A.C.E. Reduction. When developing the computer specifications for the A.C.E. cluster reduction and large block cluster subsampling, the cluster reduction design had not been determined. Thus, to accommodate several potential reduction design plans, we specified 19 strata, but only used eight.

Large Block Cluster Subsampling Input File Layout

Variable Description	Name	Pos
State	ST	1-2
A.C.E. reduction stratum	ARST	4-5
Target housing unit sample size	T	7-14

Sample Design File

The Sample Design File contains one record per block cluster selected during the listing sample selection. If the block cluster falls out of sample during the second step of the listing sample, the A.C.E. reduction, small block cluster subsampling, or the A.C.E. reduction, the remaining variables will be left blank. The initial version of the file, which will be created following the initial block cluster selection, is called SDF.US1. For each subsequent update to the file, the version number will increase by one (i.e. SDF.US2, SDF.US3). The layout for the Sample Design File is as follows:

Variable Description	Name	Places	Source
Census Region	REGION	1	UN
Census Division	DIV	2	UN
State code	STATE	3-4	UN
County code	COUNTY	5-7	UN
Local census office	LCO	8-11	CS
Interim Tract (Pseudo Tract)	ITRACT	12-17	BC
Current Sample Indicator	CSI	19	UO
A.C.E. block cluster number	CLUST	21-25	CS
Check Digit	DIGIT	26	CS
Geography block cluster number	GCLUST	28-32	BC
List/Enumerate Indicator	LEIND	33	BC
Type of Enumeration Area Recode	TEACR	34	CS
Type of Enumeration Area group	TEAG	36	BC
Number of HUs used for sample design	NHU	37-41	BC
Number of MAF HUs	NHUM	43-47	BC
Number of 1990 HUs	NHU90	49-53	BC
Sampling Stratum 1 = Small 2 = Medium 3 = Large 4 = American Indian Reservation	SS	55	UN
American Indian Country Indicator 0 = No American Indian Country 1 = American Indian Reservation/trust land 2 = Tribal Jurisdiction Area/ Alaska Native Village Statistical Area/ Tribal Designated Statistical Area	AICIND	56	BC
Demographic/Tenure Group code	DTCODE	57-58	UN
Demographic/Tenure Group label	DTLABEL	59-60	UN
Estimated Urbanicity of block cluster 1 = Urban Area with population ≥250,000 2 = Other Urban Area 3 = Non-Urban Area	ECLUSURB	62	UN
Size Category 1=Small (0-2 hus) 2=Medium (3-79 hus) 3=Large (80+ hus)	SIZCAT	63	UN
Additional space		64-91	

Variable Description	Name	Places	Source
First step index number	INDEX1	92-99	CS
Listing sample selection indicator	BC1	101	CS
1 = Selected			
Random Start for listing sample selection	RS1	103-113	UN
Take-every for listing sample selection	TE1	115-125	UN
Second step listing sample selection indicator	BC2	127	CS
0 = Not Selected			
1 = Selected			
Random Start for the second step of the listing sampling	RS2	129-139	CS
Take-every for the second step of the listing sampling	TE2	141-151	CS
Unbiased weight after block cluster sampling	WEIGHTBC	153-164	CS
Additional space		165-175	
Preliminary Number of HUs on the Independent List	NHUILP	176-180	AR
Number of Housing Units On the January 2000 DMAF	NHUDMAF	182-186	AR
Demographic Code	DEMCODE	188	AR
1 = Minority			
2 = Non-Minority			
3 = Puerto-Rico			
Consistency Code	CONCODE	189	AR
1 = Low Inconsistent (IL significantly smaller than DMAF)			
2 = Consistent			
3 = High Inconsistent ((IL significantly larger than DMAF)			
A.C.E. Reduction Stratum	ARST	190-191	AR
A.C.E. Reduction Indicator	ACERED	193	AR
0 = Not Selected			
1 = Selected			
Random Start for A.C.E. Reduction	RSAR	195-205	AR
Take-every for A.C.E. Reduction	TEAR	207-217	AR
Unbiased weight after A.C.E. reduction	WEIGHTAR	219-230	AR
Collapsing Flag	COLFLAG	232	AR
A.C.E. Reduction Index Number	INDEXR	234-241	AR
Number of Housing Units On the December 1999 DMAF (Initial)	NHUDMAFI	243-247	AR
Additional space		248-300	
Number of HUs on the Independent List	NHUIL	301-305	SB
Small Block Cluster Subsampling Stratum	SBCSS	306-307	SB
Small Block Subsampling Indicator	SB	308	SB
0 = Not Selected			
1 = Selected			
Random Start for Small Block subsampling	RSSB	310-320	SB
Initial take-every for Small Block subsampling	ITESB	322-332	SB
Unbiased weight for A.C.E. cluster	WEIGHTC	334-345	SB
Larger of the DMAF and IL HU count	LARGERHU	347-351	SB
Final take-every for Small Block subsampling	FTESB	352-362	SB
Additional space		363-370	

Variable Description	Name	Places	Source
Relisted Block Cluster Flag 0 = Not Relisted, 1 = Relisted	RELIST	371	LB
Number of total hus in block cluster	NHUEL	373-377	LB
Number of A.C.E. hus in cluster	NHUELA	379-383	LB
Number of supplemental hus in cluster	NHUELN	385-389	LB
Large Block Cluster EL subsampling code 1 = NHUELI< 80 hus, 2 = NHUELI ≥ 80 hus	ELLBSUB	391	LB
Random Start for Large Block subsampling	RSLB	393-403	LB
Take-every for Large Block subsampling	TELB	405-415	LB
Number of segments in block cluster	NSEG	417-418	LB
Number of segments selected in block cluster	NSEGSAM	420-421	LB
Day of Arrival	DAY	423-424	LB
Final Cluster Order Number	CON	431-434	LB
Number of total hus for interview in block cluster	NINT	436-440	LB
Unbiased weight for P-sample HUs	WEIGHTP	442-453	LB
Number of Assignments in block cluster	NA	455-456	LB
Final Sampling Strata	FSS	458-464	LB
Additional space		465-490	

Source Codes

AR: A.C.E. Reduction
BC: Block Clustering
CS: Block Cluster Sampling
LB: Large Block Subsampling
SB: Small Block Subsampling
UN: Universe File Creation
UO: Updated for each operation

Large Block Cluster Subsampling Parameter File Layout

Variable Description	Name	Pos
State	ST	1-2
A.C.E. reduction stratum	ARST	4-5
Target housing unit sample size	T	7-14
Number of housing units in block clusters with 80 or more housing units on the independent list	NILHUL	16-21
Number of housing units in block clusters with 0-79 housing units on the independent list (except smalls with 0-9)	NILHUM	23-28
Number of housing units in all block clusters on the independent list	NILHUT	30-35
Number of housing units in small block clusters with 0-9 housing units on the independent list	NILHUS	37-42
Take-every for the segment subsampling	TELB	44-54
Number of segments in a block cluster	NSEG	56-57
Flag for formula used for calculating NSEG	FORMULA	59
Random Number between 0 and 1	RN	61-72
Random Start for the segment subsampling	RS	74-84
Current Daily Start	DS	86-96
Cumulative Cluster Count	CCC	98-100
Daily Start for Day 1	DS1	102-112
Daily Start for Day 2	DS2	114-124
.	.	.
.	.	.
.	.	.
Daily Start for Day 20[1]	DS20	

[1]The number of days for sampling may be over or under 20. If this is the case, appropriate modifications will be made.

UNITED STATES DEPARTMENT OF COMMERCE
Bureau of the Census
Washington, DC 20233-0001

March 8, 2000

MASTER FILE

DSSD CENSUS 2000 PROCEDURES AND OPERATIONS MEMORANDUM SERIES R-26

MEMORANDUM FOR Maureen Lynch
Assistant Division Chief, Coverage Measurement Processing
Decennial Statistical Studies Division

From: Donna Kostanich DK
Assistant Division Chief, Sampling and Estimation
Decennial Statistical Studies Division

Prepared by: Ryan Cromar RC
Sample Design Team
Decennial Statistical Studies Division

Subject: Accuracy and Coverage Evaluation: Large Block Cluster Subsampling Parameter File Specifications

I. INTRODUCTION

This memorandum provides specifications for the large block cluster subsampling parameter file for the Census 2000 Accuracy and Coverage Evaluation (A.C.E.) survey. As the final stage of the A.C.E. design, large block cluster subsampling involves selecting a portion of a block cluster that has 80 or more A.C.E. housing units to be in the A.C.E. interview sample. This will be accomplished by forming segments of adjacent housing units within the block cluster and selecting a subsample of segments. The objective of large block cluster subsampling is to meet the target A.C.E. interviewing sample sizes using the most up-to-date A.C.E. housing unit counts available at the time of subsampling. The creation of the parameter file is the first step of large block cluster subsampling. The parameter file is an input for the large block cluster subsampling process documented in reference 1. The large block cluster subsampling parameter file specification is similar to the specifications prepared for the 1998 Census 2000 Dress Rehearsal and documented in reference 2.

Earlier stages of the A.C.E. sample design include the selection of A.C.E. block clusters for the listing sample (see reference 3), the A.C.E. reduction (see reference 4), and the subsampling of small block clusters (see reference 5). After the listing sample selection, the independent list is created and the results keyed and verified. Based on the results of this listing, the A.C.E. sample reduction and small block cluster subsampling are done. Subsequently, the housing unit matching and follow-up are done, and the preliminary enhanced list is created and sent to large block cluster subsampling. The preliminary enhanced list is both the input and output file for large block cluster subsampling. The output preliminary enhanced list is updated with the subsampling results, and is referred to as the subsampled preliminary enhanced list. The enhanced List is created by extracting only the housing units designated for interview following large block cluster subsampling from the subsampled preliminary enhanced list (see reference 6).

This memorandum is organized into the following sections:

- Assumptions
- Definitions
- Input
- Process
- Output
- Verification
- References

These specifications should be used to flowchart the process, to generate further discussion on requirements, to identify and finalize the record layouts of input and output files, and to write computer software to implement the methodology. During and after a testing phase, it is possible that changes to the specifications will be necessary.

If there are any questions or comments, please contact Ryan Cromar (301-457-1636), James Farber (301-457-4282), or Deborah Fenstermaker (301-457-4195) of the Decennial Statistical Studies Division (DSSD).

II. ASSUMPTIONS

The assumptions required to create the large block cluster subsampling parameter file are:

A. The creation of the large block cluster subsampling parameter file is not affected by the results of housing unit matching and follow-up. Thus the creation of this parameter file can be done using independent list housing unit counts at any time after small block cluster subsampling even if matching and follow-up are not completed. It is possible that the housing unit counts determined in this specification will differ due to housing unit follow-up.

B. Block clusters eligible for large block cluster subsampling include those that were selected in the A.C.E. reduction and remain in the sample following small block cluster subsampling. All other block clusters are not eligible for large block cluster subsampling.

C. Large block cluster subsampling is done on a flow basis over a span of several days. Therefore, the large block cluster subsampling parameter file includes fields for daily statistics that will be filled during the subsampling process and will be used to track the day-to-day results of large block cluster subsampling.

D. The large block cluster subsampling parameter file will not be revised to account for relisted block clusters. Relisted block clusters will receive the original subsampling parameters computed from the independent list before relisting. Block clusters which require relisting will not be identified nor will the relisting be done by the time the large block cluster subsampling parameters are calculated.

E. The A.C.E. housing units on the independent list are keyed and valid.

F. The A.C.E. housing units that have a Unit Status of Future Construction are excluded from the take-every calculation. Including such units could cause the A.C.E. interview sample size to be lower than expected, which would increase the variance of the A.C.E. population estimates. Excluding these units is a conservative approach to ensure that target sample sizes are achieved.

G. All decimal numbers are rounded to six digits at the time of creation using the standard rounding procedure except when noted otherwise. Decimal numbers with a seventh decimal place of five or more are rounded up in the sixth decimal place. Those with four or less in the seventh decimal place are rounded down in the sixth decimal place.

H. Note that medium and small block clusters are eligible for large block cluster subsampling since the decision to subsample is based only on the number of A.C.E. housing units in the block cluster.

III. DEFINITIONS

A. American Indian Reservation (AIR) Block Cluster

A block cluster with three or more housing units based on information available at the time block clusters were formed that is at least partially in an AIR. See Sampling Strata.

B. Block Cluster

A geographically contiguous group of Census 2000 collection blocks (see reference 7).

C. Listing Sample

The initial sampling stage of A.C.E. in which block clusters are selected for independent listing (see reference 3).

D. A.C.E. Independent List

List of all housing units in A.C.E. listing sample block clusters. The independent list is created independently of the Decennial Master Address File, the address list used for the census.

E. Keyed and Valid Housing Units

Housing units with any of the following Unit Status codes on the independent list:

1 = Occupied or vacant and intended for occupancy
2 = Under construction
3 = Future construction
4 = Unfit for habitation
5 = Boarded up
6 = Storage of household goods
7 = Vacant mobile home site
8 = Other

All of these units are included because it is possible that the unit status may change between listing and interviewing. Group quarters are not listed in A.C.E. Note that A.C.E. Future construction housing units (Unit Status = 3) are excluded from the take-every calculation as explained in section II above.

F. Large Block Cluster

See Sampling Strata.

G. Medium Block Cluster

See Sampling Strata.

H. A.C.E. Housing Unit

A housing unit on the preliminary enhanced list that is keyed and valid and has one of the following After Follow-up Match Codes: M, MU, UI, or CI.

I. A.C.E. Reduction

The process of reducing the A.C.E. listing sample from the Integrated Coverage Measurement (ICM) sample to the A.C.E. interview sample. In the A.C.E. reduction, the listing sample block clusters are subsampled, and the selected block clusters continue to small block cluster subsampling (see Reference 4).

J. A.C.E. Reduction Strata

A partition (mutually exclusive and exhaustive set) of all block clusters in a state into groups according to certain characteristics. See Attachment A for a list of the A.C.E. Reduction Strata and see reference 4 for more information on how A.C.E. reduction strata are defined.

K. Sampling Strata

A partition of all block clusters within a state into groups according to the number of housing units estimated in each cluster at the time of block clustering (see reference 7). Block Clusters are assigned to sampling strata prior to listing sample selection. The sampling strata are:

1	=	small block clusters with 0 - 2 estimated housing units
2	=	medium non-AIR block clusters with 3 - 79 estimated housing units
3	=	large non-AIR block clusters with $\geq$ 80 estimated housing units
4	=	medium and large AIR block clusters with $\geq$ 3 estimated housing units

L. Small Block Cluster

See Sampling Strata.

M. State

The 50 United States plus the District of Columbia and Puerto Rico.

IV. INPUT FILES

The inputs for the subsampling process are the following:

A. Large Block Cluster Subsampling Input File

Description:	This file contains the target housing unit sample size for each A.C.E. reduction stratum within each state.
Level:	A.C.E. reduction stratum
Scope:	One record per A.C.E. reduction stratum within each state
Layout:	See Attachment B

B. Cluster Status File

Description:	This file has one record for each of the 29,695 block clusters selected in the A.C.E. listing sample. It is updated with information from other processing stages. For large block cluster subsampling, this file is used to determine the sampling parameters.
Level:	A.C.E. Block Cluster
Scope:	All block clusters selected for the A.C.E. listing sample

C. A.C.E. Sample Design File (Version 3)

Description:	This file reflects the previous A.C.E. sampling operations: listing sample selection, A.C.E. reduction, and small block cluster subsampling.
Level:	Block Cluster
Scope:	One record for each block cluster in the A.C.E. listing sample
File Layout:	See Attachment C

V. PROCESS

Using results from the independent list, subsampling parameters for each A.C.E. reduction stratum within each state are calculated prior to subsampling any block cluster. The sampling parameters to calculate are the take-every, the target number of segments in a block cluster, and the random start. All block clusters in a common A.C.E. reduction stratum within each state have the same parameters.

A. Determine the take-every by computing the number of keyed and valid housing units on the independent list in all block clusters with 80 or more A.C.E. housing units and dividing it by the target sample size from the block clusters with 80 or more housing units, but excluding the number of future construction housing units from the calculation. Block clusters from the small sampling stratum with fewer than 10 housing units on the independent list are not included when calculating housing unit totals because these housing units are not included in the target sample size.

Use the following steps to determine the take-every for each A.C.E. reduction stratum with each state:

1. Obtain four housing unit counts for each A.C.E. reduction stratum within each state from the cluster status file. Exclude Future Construction housing units, those with Unit Status = 3, from all of these counts.

    a. The number of housing units in all block clusters from the independent list, NILHUT.

    b. The number of housing units in block clusters with 80 or more housing units from the independent list, NILHUL.

    c. The number of housing units in small block clusters (sampling stratum = 1) with fewer than ten housing units from the independent list, NILHUS.

    d. The number of housing units in block clusters with fewer than 80 housing units from the independent list excluding block clusters from step c above, NILHUM.

    e. Calculate Z to check the counts above. This is calculated by subtracting NILHUL, NILHUS and NILHUM from NILHUT.

    $$Z = NILHUT - (NILHUL + NILHUS + NILHUM)$$

    If the counts are correct, Z will be equal to zero. Resolve the cases where Z is not equal to zero.

2. Obtain the target number of sample housing units, T, for the A.C.E. reduction stratum within state from the large block cluster subsampling input file ACE2000_LBINPUT.FIN.

3. Calculate the take-every, TELB:

$$TELB = \frac{NILHUL}{T - NILHUM}$$

- If TELB < 1.012660[1], then set TELB = 1.000000.
- If T - NILHUM = 0, then contact the Sample Design Team of the DSSD.

4. Round the take-every to six decimal places.

B. Calculate the integer number of segments to be formed in a block cluster, NSEG, using the formulas below. These formulas are based on the TELB to ensure at least one segment is selected from each block cluster eligible for subsampling. In addition, create a variable called FORMULA, that codes which formula was used for calculating the number of segments.

- If TELB ≥ 2, then
  NSEG = TELB (If not an integer, round up to the next integer.)
  FORMULA = 1

- If 1 < TELB < 2, then

$$NSEG = \frac{1}{1 - \frac{1}{TELB}}$$ (If not an integer, round up to the next integer.)

  FORMULA = 2

- If TELB = 1, then
  NSEG = 1
  FORMULA = 3

[1]This value for TELB was determined to prevent any block clusters from having over 80 segments.

C. Calculate the random start, RS, by generating a random number, RN, between zero and one, rounding it to six decimal places, and multiplying it by the TELB. Round the resulting random start to six decimal places. Calculate a new random start for each A.C.E. reduction stratum within each state.

- $RS = RN \times TELB$, where $0 < RN \leq 1$
- If TELB = 1.000000, then set RS = 1.000000.

D. Starting with the large block cluster subsampling input file as a basis, create a large block cluster sampling parameter file by appending the variables created in this section plus two other variables described in step 1 below. This large block cluster subsampling parameter file will be a daily input to the large block cluster subsampling. Additional variables will be appended during future steps in the large block cluster subsampling process. This file has one record for each A.C.E. reduction stratum within each state.

1. Create two variables, current daily start, DS, and cumulative cluster count, CCC, which are needed for implementing the subsampling over several days. Initialize these variables by setting DS equal to the random start, RS, and assigning a value of zero to CCC for each A.C.E. reduction stratum within each state.

2. Update the following variables on the large block cluster subsampling parameter file using the layout in Attachment D:

   Number of housing units in block clusters with 80 or more housing units on the independent list, NILHUL
   Number of housing units in block clusters with 0-79 housing units on the independent list (except smalls with 0-9), NILHUM
   Number of housing units in all block clusters on the independent list, NILHUT
   Number of housing units in small block clusters with 0-9 housing units on the independent list, NILHUS
   Take-every for the segment subsampling, TELB
   Number of segments in a block cluster, NSEG
   Flag for formula used for calculating NSEG, FORMULA
   Random Number between 0 and 1, RN
   Random Start for the segment subsampling, RS
   Current daily start for the segment subsampling, DS
   Cumulative Cluster Number, CCC

   Round RS, DS and TELB to six decimal places. The other variables are integers.

E. As soon as the parameter file is created, provide it to the Sample Design Team in the DSSD for review and approval prior to large block cluster subsampling.

VI. OUTPUT FILE

The output requested by the Sample Design Team in the DSSD from the Process section is the following:

A. The Large Block Cluster Subsampling Parameter File

Description:	This file contains information needed for selecting the systematic subsample on a flow basis. The file will be produced after the sampling parameters are calculated. The final version will be created when the large block cluster subsampling process is complete.
Level:	A.C.E. Reduction Stratum
Scope:	One record per A.C.E. reduction stratum within each state
File Layout:	See Attachment D

VII. VERIFICATION

The following information should be provided for verification:

- Large Block Cluster Subsampling Parameter File
- Cluster Status File

Provide the sampling parameter file. See section VI above for information about this file. Access to the cluster status file is also required for verification.

VIII. REFERENCES

1 DSSD Census 2000 Procedures and Operations Memorandum Series R-27, "Census 2000 Accuracy and Coverage Evaluation: Large Block Cluster Subsampling Specifications," March 8, 2000.

2 DSSD Census 2000 Dress Rehearsal Memorandum Series A-9, "Census 2000 Dress Rehearsal ICM Sampling: Large Block Subsampling Specification," April 15, 1998.

3 DSSD Census 2000 Procedures and Operations Memorandum Series R-3, "Accuracy and Coverage Evaluation (ACE) Survey: Block Cluster Sample Selection Specification," March 29, 1999.

4 DSSD Census 2000 Procedures and Operations Memorandum Series R-, "Accuracy and Coverage Evaluation Survey: Reduction Specification," January 10, 2000, DRAFT.

5 DSSD Census 2000 Procedures and Operations Memorandum Series R-24, "Accuracy and Coverage Evaluation Survey: Small Block Cluster Subsampling," February 1, 1999.

6 DSSD Census 2000 Procedures and Operations Memorandum Series, Chapter S-HU-08, "Creation of the Census 2000 Accuracy and Coverage Evaluation (A.C.E.) Enhanced List for Person Phase Interviewing," June 21, 1999, DRAFT.

7 DSSD Census 2000 Procedures and Operations Memorandum Series R-8, "Census 2000 Specifications for Block Cluster Formation-Reissue," May 3, 1999.

cc: DSSD Census 2000 Procedures and Operations Memorandum Series Distribution List
A.C.E. Implementation Team Leaders
Statistical Design Team Leaders
Sample Design Team

A.C.E. Reduction Strata

Stratum Code[1]	Stratum Name
01	Minority
02	Non-minority Low Inconsistent
03	Non-minority Consistent
04	Non-minority High Inconsistent
05	Non-minority Inconsistent
06	Non-minority
07	Low Inconsistent
08	Consistent
09	High Inconsistent
10	Inconsistent
11	Minority Inconsistent
12	Minority Consistent
13	Full Collapse
14	Minority Low Inconsistent
15	Minority High Inconsistent
16	Medium Stratum Jumpers
17	American Indian Reservations
18	Puerto Rico
19	Small Stratum Jumpers

[1] Only Strata 01, 02, 03, 04, 16, 17, 18, and 19 were actually used for the A.C.E. Reduction. When developing the computer specifications for the A.C.E. cluster reduction and large block cluster subsampling, the cluster reduction design had not been determined. Thus, to accommodate several potential reduction design plans, we specified 19 strata, but only used eight.

Large Block Cluster Subsampling Input File Layout

Variable Description	Name	Pos
State	ST	1-2
A.C.E. reduction stratum	ARST	4-5
Target housing unit sample size	T	7-14

Sample Design File

The Sample Design File contains one record per block cluster selected during the listing sample selection. If the block cluster falls out of sample during the second step of the listing sample, the A.C.E. reduction, small block cluster subsampling, or the A.C.E. reduction, the remaining variables will be left blank. The initial version of the file, which will be created following the initial block cluster selection, is called SDF.US1. For each subsequent update to the file, the version number will increase by one (i.e. SDF.US2, SDF.US3). The layout for the Sample Design File is as follows:

Variable Description	Name	Places	Source
Census Region	REGION	1	UN
Census Division	DIV	2	UN
State code	STATE	3-4	UN
County code	COUNTY	5-7	UN
Local census office	LCO	8-11	CS
Interim Tract (Pseudo Tract)	ITRACT	12-17	BC
Current Sample Indicator	CSI	19	UO
A.C.E. block cluster number	CLUST	21-25	CS
Check Digit	DIGIT	26	CS
Geography block cluster number	GCLUST	28-32	BC
List/Enumerate Indicator	LEIND	33	BC
Type of Enumeration Area Recode	TEACR	34	CS
Type of Enumeration Area group	TEAG	36	BC
Number of HUs used for sample design	NHU	37-41	BC
Number of MAF HUs	NHUM	43-47	BC
Number of 1990 HUs	NHU90	49-53	BC
Sampling Stratum 1 = Small 2 = Medium 3 = Large 4 = American Indian Reservation	SS	55	UN
American Indian Country Indicator 0 = No American Indian Country 1 = American Indian Reservation/trust land 2 = Tribal Jurisdiction Area/ Alaska Native Village Statistical Area/ Tribal Designated Statistical Area	AICIND	56	BC
Demographic/Tenure Group code	DTCODE	57-58	UN
Demographic/Tenure Group label	DTLABEL	59-60	UN
Estimated Urbanicity of block cluster 1 = Urban Area with population ≥250,000 2 = Other Urban Area 3 = Non-Urban Area	ECLUSURB	62	UN
Size Category 1=Small (0-2 hus) 2=Medium (3-79 hus) 3=Large (80+ hus)	SIZCAT	63	UN
Additional space		64-91	

Variable Description	Name	Places	Source
First step index number	INDEX1	92-99	CS
Listing sample selection indicator	BC1	101	CS
1 = Selected			
Random Start for listing sample selection	RS1	103-113	UN
Take-every for listing sample selection	TE1	115-125	UN
Second step listing sample selection indicator	BC2	127	CS
0 = Not Selected			
1 = Selected			
Random Start for the second step of the listing sampling	RS2	129-139	CS
Take-every for the second step of the listing sampling	TE2	141-151	CS
Unbiased weight after block cluster sampling	WEIGHTBC	153-164	CS
Additional space		165-175	
Preliminary Number of HUs on the Independent List	NHUILP	176-180	AR
Number of Housing Units On the January 2000 DMAF	NHUDMAF	182-186	AR
Demographic Code	DEMCODE	188	AR
1 = Minority			
2 = Non-Minority			
3 = Puerto-Rico			
Consistency Code	CONCODE	189	AR
1 = Low Inconsistent (IL significantly smaller than DMAF)			
2 = Consistent			
3 = High Inconsistent ((IL significantly larger than DMAF)			
A.C.E. Reduction Stratum	ARST	190-191	AR
A.C.E. Reduction Indicator	ACERED	193	AR
0 = Not Selected			
1 = Selected			
Random Start for A.C.E. Reduction	RSAR	195-205	AR
Take-every for A.C.E. Reduction	TEAR	207-217	AR
Unbiased weight after A.C.E. reduction	WEIGHTAR	219-230	AR
Collapsing Flag	COLFLAG	232	AR
A.C.E. Reduction Index Number	INDEXR	234-241	AR
Number of Housing Units On the December 1999 DMAF (Initial)	NHUDMAFI	243-247	AR
Additional space		248-300	
Number of HUs on the Independent List	NHUIL	301-305	SB
Small Block Cluster Subsampling Stratum	SBCSS	306-307	SB
Small Block Subsampling Indicator	SB	308	SB
0 = Not Selected			
1 = Selected			
Random Start for Small Block subsampling	RSSB	310-320	SB
Initial take-every for Small Block subsampling	ITESB	322-332	SB
Unbiased weight for A.C.E. cluster	WEIGHTC	334-345	SB
Larger of the DMAF and IL HU count	LARGERHU	347-351	SB
Final take-every for Small Block subsampling	FTESB	352-362	SB
Additional space		363-370	

Variable Description	Name	Places	Source
Relisted Block Cluster Flag 0 = Not Relisted, 1 = Relisted	RELIST	371	LB
Number of total hus in block cluster	NHUEL	373-377	LB
Number of A.C.E. hus in cluster	NHUELA	379-383	LB
Number of supplemental hus in cluster	NHUELN	385-389	LB
Large Block Cluster EL subsampling code 1 = NHUELI< 80 hus, 2 = NHUELI ≥ 80 hus	ELLBSUB	391	LB
Random Start for Large Block subsampling	RSLB	393-403	LB
Take-every for Large Block subsampling	TELB	405-415	LB
Number of segments in block cluster	NSEG	417-418	LB
Number of segments selected in block cluster	NSEGSAM	420-421	LB
Day of Arrival	DAY	423-424	LB
Final Cluster Order Number	CON	431-434	LB
Number of total hus for interview in block cluster	NINT	436-440	LB
Unbiased weight for P-sample HUs	WEIGHTP	442-453	LB
Number of Assignments in block cluster	NA	455-456	LB
Final Sampling Strata	FSS	458-464	LB
Additional space		465-490	

Source Codes

AR: A.C.E. Reduction
BC: Block Clustering
CS: Block Cluster Sampling
LB: Large Block Subsampling
SB: Small Block Subsampling
UN: Universe File Creation
UO: Updated for each operation

Large Block Cluster Subsampling Parameter File Layout

Variable Description	Name	Pos
State	ST	1-2
A.C.E. reduction stratum	ARST	4-5
Target housing unit sample size	T	7-14
Number of housing units in block clusters with 80 or more housing units on the independent list	NILHUL	16-21
Number of housing units in block clusters with 0-79 housing units on the independent list (except smalls with 0-9)	NILHUM	23-28
Number of housing units in all block clusters on the independent list	NILHUT	30-35
Number of housing units in small block clusters with 0-9 housing units on the independent list	NILHUS	37-42
Take-every for the segment subsampling	TELB	44-54
Number of segments in a block cluster	NSEG	56-57
Flag for formula used for calculating NSEG	FORMULA	59
Random Number between 0 and 1	RN	61-72
Random Start for the segment subsampling	RS	74-84
Current Daily Start	DS	86-96
Cumulative Cluster Count	CCC	98-100
Daily Start for Day 1	DS1	102-112
Daily Start for Day 2	DS2	114-124
.	.	.
.	.	.
.	.	.
Daily Start for Day 20[1]	DS20	

[1] The number of days for sampling may be over or under 20. If this is the case, appropriate modifications will be made.

UNITED STATES DEPARTMENT OF COMMERCE
Bureau of the Census
Washington, DC 20233-0001

MASTER FILE

March 8, 2000

DSSD CENSUS 2000 PROCEDURES AND OPERATIONS MEMORANDUM SERIES R-27

MEMORANDUM FOR Maureen Lynch
Assistant Division Chief, Coverage Measurement Processing
Decennial Statistical Studies Division

From: Donna Kostanich DK
Assistant Division Chief, Sampling and Estimation
Decennial Statistical Studies Division

Prepared by: Ryan Cromar RC
Sample Design Team

Subject: Accuracy and Coverage Evaluation: Large Block Cluster Subsampling Specifications

I. INTRODUCTION

This memorandum provides specifications for large block cluster subsampling for the Census 2000 Accuracy and Coverage Evaluation (A.C.E.) survey. As the final stage of the A.C.E. sample design, large block cluster subsampling involves selecting a portion of a block cluster that has 80 or more A.C.E. housing units (HUs) to be in the A.C.E. interview sample. This will be accomplished by forming segments of adjacent HUs within the block cluster and selecting a subsample of segments. The objective of large block cluster subsampling is to meet the target A.C.E. interviewing sample sizes using the most up-to-date A.C.E. HU counts available at the time of subsampling. The large block cluster subsampling specification is similar to the specifications prepared for the 1998 Census 2000 Dress Rehearsal and documented in reference 1. The creation of the initial large block cluster subsampling parameter file, which is required to conduct large block cluster subsampling, is specified in reference 2.

These specifications also include instructions for assigning supplemental HUs to segments. Supplemental HUs are units found only in the census address list and not in the A.C.E. independent address list, and are not eligible for the A.C.E. interview sample. However, they must go through large block subsampling so that they are properly prepared for E-Sample Identification. A third task included in these specifications is the creation of interviewer workload assignments when the total number of HUs to interview in a block cluster exceeds 80.

Earlier stages of the A.C.E. sample design include the selection of A.C.E. block clusters for the listing sample (see reference 3), the A.C.E. reduction (see reference 4), and the subsampling of small block clusters (see reference 5). After the listing sample selection, the independent listing is completed, the results keyed and verified, and the Independent List (IL) is created. Based on the results of this listing, the A.C.E. sample reduction and small block cluster subsampling are done. Subsequently, the HU matching and follow-up operations are done, and the preliminary Enhanced List (EL) is created and sent to large block cluster subsampling. The preliminary EL is both the input and output file for large block cluster subsampling. The output preliminary EL is updated with the results of subsampling, and is referred to as the subsampled preliminary EL. The Enhanced List is created by extracting only housing units designated for interview following large block cluster subsampling from the subsampled preliminary EL (see reference 6).

This memorandum is organized into the following sections:

- Assumptions
- Definitions
- Process Overview
- Input
- Process
- Output
- Verification
- References

These specifications should be used to flowchart the process, to generate further discussion on requirements, to identify and finalize the record layouts of input and output files, and to write computer software to implement the methodology. During and after a testing phase, it is possible that changes to the specifications will be necessary.

If there are any questions or comments, please contact Ryan Cromar (301-457-1636), James Farber (301-457-4282), or Deborah Fenstermaker (301-457-4195) of the Decennial Statistical Studies Division (DSSD).

II. ASSUMPTIONS

The assumptions required for these specifications are:

A. Block clusters eligible for large block cluster subsampling include those that were selected in the A.C.E. reduction and remain in the sample following small block cluster subsampling. All other block clusters are not eligible for large block cluster subsampling.

B. For each block cluster in the listing sample, matching and HU follow-up operations have been completed and After Follow-up Match Codes have been assigned. Note that List/Enumerate and Relisted block clusters will not go through matching and follow-up but will have After Follow-up Match Codes assigned.

C. The A.C.E. HUs are the only HUs that are eligible for the A.C.E. interview sample. A.C.E. HUs are defined as HUs from the IL which have survived all HU follow-up procedures. These units have a match code of M, MU, UI, or CI. Refer to the definitions of the codes in section III below. Note that even though future construction was not included in the calculations of the sampling parameters, any future construction addresses that survived HU follow-up and matching are included as A.C.E. HUs.

D. Note that medium and small block clusters are eligible for large block cluster subsampling since the decision to subsample is based only on the number of A.C.E. housing units in the block cluster.

E. All A.C.E. HUs on the IL are keyed and valid.

F. All decimal numbers are rounded to six digits at the time of creation using the standard rounding procedure except when noted otherwise. Decimal numbers with a seventh decimal place of five or more are rounded up in the sixth decimal place. Those with four or less in the seventh decimal place are rounded down in the sixth decimal place.

G. The initial large block cluster subsampling parameter file has been created and verified.

H. Large block cluster subsampling is done on a flow basis over a span of several days. Therefore, daily large block cluster subsampling parameter files will be created and will include daily statistics to track the day-to-day results of large block cluster subsampling and will provide the needed inputs for processing on successive days. Large block cluster subsampling may run concurrently with operations such as HU follow-up.

I. There will be no large block cluster subsampling in American Indian Reservations.

III. DEFINITIONS

A. After Follow-up Match Code

Code assigned to HUs after HU Follow-up. For the purposes of these specifications, the only match codes that need to be defined are those that occur on the preliminary EL. As documented in references 6 and 7, these match codes are:

- M = The A.C.E. and census addresses match.
- MU = The A.C.E. and census addresses match and there is not enough information on the follow-up form to confirm this match as a HU with certainty. The follow-up interview was not done, was incomplete, was never sent, had contradictory information, or was a noninterview.
- UI = Not enough information on the follow-up form to assign a code to the nonmatched A.C.E. HU with certainty. The follow-up interview was not done, was incomplete, was never sent, had contradictory information, or was a noninterview.
- UE = Not enough information on the follow-up form to assign a code to the census nonmatched HU with certainty. The follow-up interview was not done, was incomplete, was never sent, had contradictory information, or was a noninterview.
- CI = The A.C.E. housing unit existed as a HU at the time of the follow-up interview and is correctly geocoded in the block cluster. The HU is not found in the census.
- CE = The census housing unit existed as a HU at the time of the follow-up interview and is correctly geocoded in the block cluster. The HU is not found in the A.C.E..

B. American Indian Reservation Block Cluster

A block cluster with three or more HUs based on information available at the time block clusters were formed that is at least partially in an American Indian Reservation (AIR). See Sampling Strata.

C. Block Cluster

A geographically contiguous group of Census 2000 collection blocks (see reference 8).

D. A.C.E. Housing Unit

A housing unit on the preliminary EL that is keyed and valid and has one of the following After Follow-up Match Codes: M, MU, UI, or CI.

E. Housing Unit Follow-up

Reconciliation in the field of HUs not matched in the Matching process.

F. A.C.E. Independent List

List of all HUs in A.C.E. listing sample block clusters. The IL is created independently of the Decennial Master Address File (DMAF), the address list used for the census.

G. Keyed and Valid HUs

A.C.E. HUs with the following Unit Status:

1 = Occupied or vacant and intended for occupancy
2 = Under construction
3 = Future construction
4 = Unfit for habitation
5 = Boarded up
6 = Storage of household goods
7 = Vacant mobile home site
8 = Other

All of these units are included because it is possible that the unit status may change between listing and interviewing. Group quarters are not listed in A.C.E.

H. Large Block Cluster

See Sampling Strata.

I. Listing Sample

The initial sampling stage of the A.C.E. survey in which block clusters are selected for independent listing (see reference 3).

J. Matching

Computer and clerical process of comparing the IL and the DMAF to determine which addresses are common to both lists and which are on only one list.

K. Medium Block Cluster

See Sampling Strata.

L. Preliminary Enhanced List

The input and output of the large block cluster subsampling process. The output is referred to as the subsampled preliminary EL.

M. A.C.E. Reduction

The process of reducing the A.C.E. listing sample from the Integrated Coverage Measurement (ICM) sample to the A.C.E. interview sample. In the A.C.E. reduction, the listing sample block clusters are subsampled, and the selected clusters continue to small block cluster subsampling (see Reference 4).

N. A.C.E. Reduction Strata

A partition (mutually exclusive and exhaustive set) of all block clusters in a state into groups according to certain characteristics. See Attachment A for a list of the A.C.E. Reduction Strata and see reference 4 for more information on how A.C.E. reduction strata are defined.

O. Sampling Strata

A partition of all block clusters within a state into groups according to the number of HUs estimated in each cluster at the time of block clustering (see reference 9). Sampling strata were assigned to block clusters prior to listing sample selection. The sampling strata are:

1	=	small block clusters with 0 - 2 estimated HUs
2	=	medium non-AIR block clusters with 3 - 79 estimated HUs
3	=	large non-AIR block clusters with ≥ 80 estimated HUs
4	=	medium and large AIR block clusters with ≥ 3 estimated HUs

Note that medium and even small block clusters are eligible for large block cluster subsampling since the decision to subsample is based only on the number of A.C.E. HUs in the block cluster.

P. Small Block Cluster

See Sampling Strata.

Q. State

The 50 United States plus the District of Columbia and Puerto Rico.

R. Supplemental Housing Unit

A housing unit found in the version of the DMAF used for housing unit matching and not in the IL. Supplemental HUs are not A.C.E. HUs and are not eligible for the A.C.E. interview sample. However, they are assigned to segments during large block cluster subsampling to facilitate E-Sample Identification. The After Follow-up Match Codes for the supplemental HUs are CE and UE.

IV. PROCESS OVERVIEW

Large block cluster subsampling is used to achieve the target A.C.E. interview sample size in each A.C.E. reduction stratum within each state. These target sample sizes are designed to increase the weights of large block clusters so as to minimize weight variation between medium and large block clusters in the same reduction stratum within a state. This overview will detail the steps of the large block cluster subsampling process. The steps listed below correspond to the steps in section VI below, which contains the programming instructions and is significantly less descriptive than this overview.

A. Create Block Cluster HU Counts and Subsampling Status Codes

The preliminary EL includes not only A.C.E. HUs but also supplemental HUs, therefore a step is required to determine the various HU counts. This step also determines if large block cluster subsampling is required. For each block cluster, the following HU counts are computed:

- total number of preliminary EL HUs
- total number of A.C.E. HUs
- total number of supplemental HUs

Using the counts, the following table is used to determine if large block cluster subsampling is required. Large block cluster subsampling occurs only in certain block clusters, as shown in Table 1:

Table 1. Where Large Block Cluster Subsampling is Required

# A.C.E. HUs	In AIR?	Lg. BC Subsampling Required?
Less than 80	Yes	No
Less than 80	No	No
80 or more	Yes	No
80 or more	No	Yes

Table 1 shows that only block clusters that have 80 or more A.C.E. HUs and are not in an AIR go through large block cluster subsampling. There is no subsampling on an AIR regardless of the size of the block cluster. Nor is there subsampling if the number of supplemental HUs exceeds a certain number. Note that sampling stratum also does not affect the need for subsampling. A block cluster that was originally designated as small or medium may end up with 80 or more A.C.E. HUs after independent listing, and therefore would go through large block cluster subsampling.

Block clusters that do not require large block cluster subsampling skip to step E below. Block clusters that do require subsampling continue to step B.

B. Create Segments

Only block clusters that have 80 or more A.C.E. HUs and that are not in an AIR need to be segmented for large block cluster subsampling. The number of segments to form in block clusters in each A.C.E. reduction stratum within each state was determined during the creation of the initial large block cluster subsampling parameter file, but the size of the segments, the number of A.C.E. HUs in each segment, will vary among block clusters in the same reduction stratum within a state. Determining the segment size for a block cluster is the first step in the creation of segments. The segment size calculation is simply the number of A.C.E. HUs divided by the prespecified number of segments. Since this segment size will usually not result in an integer sample size, an algorithm is used to distribute the remainder among segments.

After the segment sizes for a block cluster are determined, the A.C.E. HUs are assigned to segments. A.C.E. HUs are assigned to the first segment until that segment's size is reached. A.C.E. HUs are then assigned to the second segment until its target size is reached, and so on until all segments attain their designated sizes.

The final step of segment creation is the assignment of supplemental HUs to segments. Supplementals are not eligible for the A.C.E. interview sample, but they still require a segment identifier to facilitate E-Sample Identification. Supplemental HUs are assigned to the same segment as the nearest preceding A.C.E. HU. It is operationally impossible for a supplemental HU to occur before an A.C.E. HU in any block cluster that has 80 or more A.C.E. HUs on the preliminary EL, so there will always exist a preceding A.C.E. HU for all supplementals in the segmenting step.

C. Create Segment Level Variables and Codes

For each segment, total HU counts similar to those in step A above need to be computed. The HU counts required are the total number of HUs, the total number of A.C.E. HUs, and the total number of supplemental HUs. These counts are used to compute interviewing sample sizes following the selection of the subsample.

D. Select a Sample of Segments for Each A.C.E. Reduction Stratum within State

This is the actual subsampling step. A systematic sample of segments is selected within each A.C.E. reduction stratum and state for inclusion in the A.C.E. interview sample. A complication of this subsampling operation is that block clusters arrive into this step on a flow basis. Ideally, the subsampling of segments would be done as a single operation after all matching and follow-up operations have been completed and all block clusters have been placed on the preliminary EL. However, it is essential that the interview sample get to the field as quickly as possible, and thus large block cluster subsampling will be performed on a daily basis until all eligible block clusters have been processed. Despite the daily processing, the subsampling is designed as if it were a single operation, where the universe is all block clusters available on a given day instead of all block clusters in the A.C.E. sample. The only loss is the ability to sort all block clusters together. Instead, a geographic sort of block clusters for each day will be done to minimize geographic bias in the interview sample.

A sample of segments is selected from all block clusters available for subsampling on a given day and in the same A.C.E. reduction stratum and state. At the end of each day, the point where the sampling ends is carried over as the starting point for the next day so that only one random start is required for each A.C.E. reduction stratum within each state. Subsampling data from each day is saved on the large block cluster subsampling parameter file for verification purposes. The sample of segments is selected using the standard systematic sampling technique with a random start.

E. Identify Housing Units for the A.C.E. Interview Sample

After segments have been selected for the A.C.E. interview sample, the appropriate HUs in those segments need to be identified for interviewing. Only A.C.E. HUs in selected segments are in the A.C.E. interview sample. Supplemental HUs in selected segments are not designated for interview. An interview flag is set on the preliminary EL for A.C.E. HUs that are designated for interviewing. A later operation extracts the HUs to be interviewed from the subsampled preliminary EL, and the resulting file is the Enhanced List.

F. Create Interviewer Workload Assignments

This step occurs only for block clusters with 80 or more interview HUs. In these block clusters, interviewer assignments of 40 to 50 HUs will be created to facilitate interviewing efficiency. Assignment workloads will be balanced such that the maximum difference between any two assignments in a cluster will be one HU. The maximum number of assignment workload areas in a single cluster is 26.

G. Update or Create Files

Files such as the Sample Design File and the Large Block Subsampling Segment File will be updated throughout the process to facilitate subsampling verification. At the end of each day, the Large Block Cluster Subsampling Parameter File will also be updated to provide information to begin processing on the next day. The final versions of these files will be created at the end of large block cluster subsampling. Note that one of the results of large block cluster subsampling will be the final A.C.E. weight for each interviewed HU. This weight is computed as the product of the take-everys from all previous sampling operations.

This step also updates files for block clusters that have zero HUs and thus were not eligible for large block cluster subsampling.

V. INPUT

The input sources for the large block cluster subsampling process are the following:

A. Initial Large Block Cluster Subsampling Parameter File

Description:	This file contains sampling parameters needed for selecting the systematic sample on a flow basis. The final version will be created when production is complete. The creation of this file is documented in reference 2.
Level:	A.C.E. Reduction Stratum
Scope:	One record per A.C.E. reduction stratum within each state
Layout:	See Attachment A

B. Cluster Status File

Description:	This file has one record for each block cluster selected for the A.C.E. listing sample. It is updated with information from other processing stages. For large block cluster subsampling, this file is used to determine the subsampling parameters and to determine when a block cluster is available for subsampling.
Level:	Block Cluster
Scope:	One record for each block cluster in the A.C.E. listing sample

C. Daily Large Block Cluster Subsampling Parameter File

Description:	This file contains information for selecting the systematic sample on a flow basis. The file will be produced after the sampling parameters are calculated and will be updated daily during large block cluster subsampling to record the starting point for the next day's systematic sampling. The final version will be created when production is complete.
Level:	A.C.E. Reduction Stratum
Scope:	One record per A.C.E. reduction stratum within each state
Layout:	See Attachment A

D. Preliminary Enhanced List

Description:	The file is created from the matching of the IL and the DMAF, and the associated housing unit follow-up of the non-matches. The types of HUs on this file are IL only, IL and DMAF, and supplementals (DMAF only).
Level:	Housing Unit
Scope:	One record for each HU in block clusters selected for A.C.E. following small block cluster subsampling.
Layout:	See Attachment B

E. A.C.E. Sample Design file (Version 3)

Description:	This file reflects the previous A.C.E. sampling operations: listing sample selection, A.C.E. reduction, and small block cluster subsampling.
Level:	Block Cluster
Scope:	One record for each block cluster in the A.C.E. listing sample
File Layout:	See Attachment C

VI. PROCESS

The following are the steps of large block cluster subsampling. See section IV above for a detailed overview of these steps. These steps are completed on a flow basis for each block cluster that remains in the A.C.E. sample following small block cluster subsampling. For those block clusters that have fewer than 80 A.C.E. HUs, the process is simple since no segmenting or subsampling is required. The process is more complex for the block clusters that have 80 or more A.C.E. HUs. Attachment D gives an example of the process, and Attachment E provides a flowchart.

A. Create Block Cluster Housing Unit Counts and Subsampling Status Codes

Create nine variables for each block cluster regardless of cluster size to track processing.

1. Create the variable DAY to record the day on which a block cluster arrives after computer matching and HU follow-up. When the first block cluster arrives, set DAY to 1. Increment DAY by 1 on the next day until all block clusters have arrived. On a particular day, assign the same DAY value to all block clusters processed on that day.

2. Determine HU counts from the preliminary EL in each block cluster as follows:

   a. Count the total number of HUs, NHUEL.

   b. Count the total number of A.C.E. HUs, NHUELA.

   c. Count the total number of supplemental HUs, NHUELN.

   d. Calculate Z to check the counts above.

   $$Z = NHUEL - (NHUELA + NHUELN)$$

   If the counts are correct, Z will be equal to zero. Resolve the cases where Z is not equal to zero.

3. Use the HU counts calculated above and the AICIND variable from the Sample Design File for the AIR status of the block cluster to assign the subsampling status codes ELLBSUB, NSEGSAM, SEGSUB, and SEGID according to Table 2:

   - ELLBSUB is a block cluster flag to denote if subsampling is required, and is assigned to the entire block cluster.
   - NSEGSAM is the number of segments in sample in the block cluster, and is also assigned to the entire block cluster.
   - SEGSUB is the flag that indicates if a segment is in sample or not, and is assigned to individual segments within a block cluster.
   - SEGID is a two-character code identifying each segment and the HUs in each segment. SEGID is assigned to segments and HUs.

   Note that subsampling status codes are unknown at this step for block clusters with 80 or more A.C.E. HUs and not in an AIR.

   Table 2 also indicates to which step a block cluster should proceed for each possible block cluster status. Block clusters that do not require subsampling proceed to the interview sample identification. Block clusters that do require subsampling go through the segmenting and subsampling process before the interview sample can be identified.

Table 2. Summary of Subsampling Status Codes

Value of NHUELA	AICIND	ELLBSUB	NSEGSAM	SEGSUB	SEGID	Go to Step
less than 80	1	1	1	1	AA	E
less than 80	0 or 2	1	1	1	AA	E
80 or more	1	1	1	1	AA	E
80 or more	0 or 2	2	Unknown	Unknown	Unknown	B

B. Create Segments

Only block clusters with 80 or more A.C.E. HUs that are not on an AIR need to be segmented. For each block cluster that requires segmenting, do the following:

1. Calculate the number of A.C.E. HUs in each segment as follows:

   a. Determine the number of segments to form from the variable NSEG on the initial large block cluster subsampling parameter file. If NSEG = 1, assign SEGID = AA to all HUs in the cluster and proceed to step E below and treat the block cluster as if it had fewer than 80 A.C.E. HUs or was on an AIR.

   b. Compute the average number of A.C.E. HUs per segment as
   $$\frac{NHUELA}{NSEG}.$$

   c. Truncate this number to an integer, and denote the truncated value AVGSEG. Compute the number of remaining A.C.E. HUs, R, as R = NHUELA - (NSEG × AVGSEG)

   d. Assign SEGSIZE for each segment as follows:

      - For the first R segments, SEGSIZE = AVGSEG + 1
      - For the remaining NSEG - R segments, SEGSIZE = AVGSEG

2. Within each block cluster, assign A.C.E. HUs to segments as follows:

   a. Sort all preliminary EL HUs (A.C.E. and supplemental HUs) by map spot number (MSN) and within-map spot number ID (MSNID).

   b. Assign SEGID and A.C.E. HUs to the NSEG segments. The first segment in the block cluster receives SEGID AA. Assign A.C.E. HUs to this segment until it contains its predetermined number of A.C.E. HUs as indicated by its SEGSIZE. Assign SEGID values for each A.C.E. HU in the block cluster the same value as SEGID for their segment. The second segment is then given a SEGID of BA, and A.C.E. HUs are assigned to segment BA until its value of SEGSIZE is reached. Continue with segment CA and so forth in the same manner until all NSEG segments contain as many A.C.E. HUs as their values of SEGSIZE dictate. If there are more than 26 segments in a block cluster, continue SEGID as AB, BB, CB, and so forth.

3. Assign supplemental HUs to the same segment as the preceding A.C.E. HU. A supplemental HU cannot precede the first A.C.E. HU in a block cluster.

C. Create Segment Level Variables and Codes

Determine HU counts in each segment in each segmented block cluster as follows:

1. Count the total number of HUs in the segment, NHUELS.

2. Count the number of A.C.E. HUs in the segment, NHUELAS.

3. Count the number of supplemental HUs in the segment, NHUEI.NS.

4. Calculate Z to check the counts above:

   Z = NHUELS - (NHUELAS + NHUELNS)

   If the counts are correct, Z will be equal to zero. Resolve the cases where Z is not equal to zero.

D. Select a Sample of Segments within A.C.E. Reduction Stratum and State

Do the following sampling procedure separately for each A.C.E. reduction stratum within each state. Sort the block clusters available on a given day by block cluster number within A.C.E. reduction stratum within state before subsampling.

1. Create the block cluster level variable CON to identify the order in which the block clusters are arranged within a particular A.C.E. reduction stratum and state prior to sampling:

   a. Obtain the cumulative cluster count, CCC, from the large block cluster subsampling parameter file from the previous day.

   b. Set CON = CCC + 1 for the first block cluster processed on a given day. Increment CON by one for each remaining block cluster to be processed on that day.

   c. When the last block cluster on a given day has been processed, set CCC = CON for the last block cluster, and save CCC to the parameter file for the current day.

2. Create the segment level variable DSON to identify the order in which the segments are arranged within block clusters in a particular A.C.E. reduction stratum and state on a single day as follows:

   a. Set DSON = 1 for segment AA in the first block cluster in the first A.C.E. reduction stratum within state to be subsampled that day.

   b. Increment DSON by one for each remaining segment in all block clusters in that same A.C.E. reduction stratum within a state to be subsampled that day.

   c. Let N = the maximum value for DSON for the A.C.E. reduction stratum and state on that day.

3. Select a systematic subsample of segments for each state and A.C.E. reduction stratum:

   Generate a sequence of numbers $L_1, ..., L_n$ as follows:

   a. Obtain the current daily start value, DS, and the take-every for the current A.C.E. reduction stratum within each state, TELB, from the current day's parameter file.

   b. Let $L_1 = DS$.

   c. Calculate $L_j = L_{j-1} + TELB$, for $j = 2, ..., n$, where n is the largest integer such that $[DS + (n - 1) \times TELB] \leq N$.

   d. Round each L_j up to the nearest integer (an integer rounds to itself).

   e. For each segment in the reduction stratum with a DSON equal to the rounded values of L_j, $j = 1, ..., n$, assign SEGSUB = 1. These segments are in the A.C.E. interview sample.

   f. For each segment in the sampling stratum with a DSON not equal to the rounded values of L_j, $j = 1, ..., n$, assign SEGSUB = 0. These segments are not in the sample.

   g. Calculate the daily end value, $DE = DS + (TELB \times n) - N$.

   h. Save DE as DS on the next day's sampling parameter file.

   i. In addition, to monitor the day-to-day sampling progress, keep track of the daily start value. For each day of sampling, update DS on the large block cluster subsampling parameter file. Name these variables according to the day of sampling (i.e. DS1 is the daily start for day 1, DS2 is the daily start for day 2, ..., DS20 is the daily start for the final day[1]). If no sampling is conducted in an A.C.E. reduction stratum for a particular day, the corresponding variable will be blank.

---

[1] The number of days for sampling may be over or under 20. If this is the case, make the appropriate modifications.

For example:

On day one, if N = 40, TELB = 4.500000, and RS = DS = 1.553000, then n = 9. Set L_1 = 1.553000. The generated L_js would be the sequence: 1.553000, 6.053000, 10.553000, 15.053000, 19.553000, 24.053000, 28.553000, 33.053000, and 37.553000. Therefore, the segments with DSON values of 2, 7, 11, 16, 20, 25, 29, 34, and 38 would be selected for the sample. The daily end is DE = 1.553000 + 4.500000×9 - 40 = 2.053000.

On day two, if N = 15, TELB = 4.500000, and DS = 2.053000, then n = 3. Set L_1 = 2.053000. The generated L_js would be the sequence: 2.053000, 6.553000, and 11.053000. Therefore, the segments with DSON values of 3, 7, and 12 would be selected for the sample. The daily end is DE = 2.053000 + 4.500000×3 - 15 = 0.553000. This continues until all block clusters have been processed.

4. Check the number of sampled segments daily by calculating c:

$$c = \left|\frac{N}{TELB} - n\right|$$

If the sampling is implemented correctly, c will be less than 1. For values of c that are not less than one and have not been resolved, contact the Sample Design Team for review of the sampling operations.

5. For each block cluster, count the number of segments selected to remain in sample, NSEGSAM.

E. Identify Housing Units for the A.C.E. Interview Sample

All A.C.E. HUs in selected segments will be sent to interview. All A.C.E. HUs in unselected segments and supplemental HUs in all segments will not be sent to interview.

1. Note that block clusters that did not undergo subsampling rejoin the process at this point. These block clusters and their A.C.E. HUs already have SEGSUB, NSEGSAM, and SEGID values assigned. Assign the following fields for these block clusters:

   - SEGSIZE = NHUELA for the one segment in the block cluster
   - NHUELS = NHUEL for the one segment in the block cluster
   - NHUELAS = NHUELA for the one segment in the block cluster
   - NHUELNS = NHUELN for the one segment in the block cluster
   - CON = Blank for the A.C.E. HUs in the block cluster
   - DSON = Blank for the A.C.E. HUs in the block cluster
   - Assign SEGID = AA for all supplemental HUs in the block cluster

2. For all block clusters, regardless of whether or not they were subsampled, create an A.C.E. interview flag for each A.C.E. and supplemental HU, INTERVW, and assign as follows:

   - If SEGSUB = 1 then
     INTERVW = 1 for all A.C.E. HUs
     INTERVW = 9 for all supplemental HUs
   - If SEGSUB = 0, then
     INTERVW = 0 for all A.C.E. HUs
     INTERVW = 8 for all supplemental HUs

3. Compute interview HU counts for use in forming workload assignments as follows:

   a. Count the number of total HUs for interview in each block cluster, NINT.

   b. Count the number of total HUs for interview in each segment, NINTS.

F. Create Interviewer Workload Assignments

Create manageable interviewer workload assignments in block clusters with 80 or more HUs to interview.

1. Calculate the number of assignments needed for each block cluster, NA, and the size of the assignments, ASIZE, as follows:

- If NINT < 80, then
  NA = 1
  ASIZE = NINT

- If NINT >= 80, then

  a. Compute $NA = \frac{NINT}{40}$. If NA is not an integer, round it down to the next integer. If NA > 26, set NA = 26.

  b. Calculate the average number of HUs per assignment, AVGHUA, as follows:

  $$\frac{NINT}{NA}.$$

  Truncate this number to an integer. AVGHUA is the truncated value.

  c. Calculate the number of remaining HUs for interview, RINT:

  RINT = NINT - (NA × AVGHUA)

  d. Calculate ASIZE for each assignment as follows:

  - For the first RINT assignments,
    ASIZE = AVGHUA + 1

  - For the remaining NA - RINT assignments,
    ASIZE = AVGHUA

2. Assign the assignment identifier to interview HUs in each segment as follows:

Create an assignment identifier, ASSIGNID, to distinguish among assignments within a block cluster.

- If NINT < 80, then
  ASSIGNID = AA for all interview HUs in the block cluster

- If NINT >=80, then

  a. Sort the interview HUs by MSN and MSNID in the block cluster.

  b. Assign ASSIGNID and interview HUs to the NA assignments. The first assignment in the block cluster receives ASSIGNID AA. Assign interview HUs to this assignment until it contains its predetermined number of interview HUs as indicated by its ASIZE. The second assignment is then given an ASSIGNID of AB, and interview HUs are assigned to assignment AB until its value of ASIZE is reached. Continue with assignment AC and so forth in the same manner until all NA assignments contain as many interview HUs as their values of ASIZE dictate. The maximum value of ASSIGNID is AZ since no more than 26 assignments can be created in a block cluster.

  c. Use ASSIGNID to distinguish workload assignments on the preliminary EL.

G. Update or Create Files

1. Daily and Final Large Block Cluster Subsampling Parameter Files

At the end of each day, update the daily large block cluster subsampling parameter file for that day with the following information:
Daily Start, DS
Cumulative Cluster Count, CCC
Daily Start for the Current Day *i*, DS*i*

On the final day of processing, create the final large block cluster subsampling parameter file by copying the final day's parameter file.

2. Preliminary Enhanced List

Append the following HU variables to the preliminary EL during the large block cluster subsampling process:

Segment Identifier, SEGID
Assignment Identifier, ASSIGNID
Interview Flag, INTERVW

3. Daily and Final Large Block Cluster Subsampling Segment Files

Create a file each day that includes the segments created in block clusters that required more than one segment. Include the following information in these files:

Variable Description	Name	Char
State	STATE	1-2
County	CTY	4-6
A.C.E. Reduction Stratum	ARST	8-9
Sampling Stratum	SS	11
A.C.E. Block cluster number and Check Digit	CLUST	13-18
Day of Arrival	DAY	20-21
Segment Identifier	SEGID	23-24
Daily Segment Order Number	DSON	26-28
Number total HUs in segment	NHUELS	30-34
Number A.C.E. HUs in segment	NHUELAS	36-40
Number supplemental HUs in segment	NHUELNS	42-46
Segment subsampling code	SEGSUB	48
Number of HUs for interview in segment	NINTS	50-54

On the final day of processing, create a final large block cluster subsampling segment file by concatenating the segment files for all of the days into a single file.

4. Daily Sample Design Files

Update version three of the Sample Design File with the results of large block cluster subsampling for the first day and then on each successive day update the preceding day's Daily Sample Design File by appending the following block cluster level information. Create the variables RELIST and WEIGHTP for the Sample Design File:

- RELIST: Relisted block cluster flag
  - 0 = Block cluster not relisted
  - 1 = Block cluster relisted
- WEIGHTP: Block cluster A.C.E. weight
  - WEIGHTP = TELB x TE1 x TE2 x TEAR x FTESB (Last four TEs from Sample Design File)
- FSS: Final Sampling Stratum
  - Concatenate the following variables:
    - State, ST
    - Small Block Cluster Subsampling Stratum, SBCSS
    - A.C.E.Reduction Stratum, ARST

Variable Description	Name	Char
Relisted block cluster flag	RELIST	371
Number of total HUs in block cluster	NHUEL	373-377
Number of A.C.E. HUs in block cluster	NHUELA	379-383
Number of supplemental HUs in block cluster	NHUELN	385-389
Large block cluster subsampling code	ELLBSUB	391
Random Start for large block cluster subsampling	RSLB	393-403
Take-every for large block cluster subsampling	TELB	405-415
Number of segments in block cluster	NSEG	417-418
Number of segments selected in block cluster	NSEGSAM	420-421
Day of Arrival for block cluster	DAY	423-424
Final Cluster Order Number	CON	431-434
Number of total HUs for interview in block cluster	NINT	436-440
Unbiased Weight for P-sample HUs	WEIGHTP	442-453
Number of Assignments in block cluster	NA	455-456
Final Sampling Stratum	FSS	458-464

5. Sample Design File, Version 4

After the last day of large block cluster subsampling, create version 4 of the Sample Design File as follows:

a. For block clusters that went through large block subsampling, including those that did not require segmenting or subsampling, the Sample Design File will have been updated during the process so no further updates are required.

b. For block clusters in the A.C.E. sample (CSI = 1 on the Sample Design File) that have zero total HUs, assign values to variables as follows and include these fields in version 4 of the Sample Design File using the layout in section VI.G.4 above.

- RELIST, WEIGHTP, and FSS are defined as above
- Set the following fields to one: ELLBSUB, NSEG, NSEGSAM, DAY, NA, RSLB, TELB (RSLB and TELB are decimal numbers with six digits after the decimal)
- Set the following fields to zero: NINT, NHUEL, NHUELA, NHUELN
- Set CON to Blank

c. For block clusters not in the A.C.E. sample (CSI = 0 on the Sample Design File), blank all fields listed in section VI.G.4 above.

VII. OUTPUT

The outputs requested by the Sample Design Team are the following:

A. Daily and Final Large Block Cluster Subsampling Parameter Files

See section V for the description of these files and Attachment A for the layout of these files.

B. Subsampled Preliminary Enhanced List

Description: The subsampled preliminary EL is the input preliminary EL updated with the results of large block cluster subsampling. All HUs in A.C.E. sample block clusters are on the subsampled preliminary EL. The Enhanced List will be created by extracting the HUs designated for interview from the subsampled preliminary EL. The subsampled preliminary EL also provides the input for E-Sample Identification.

Level: Housing Unit

Scope: One record for each HU in block clusters selected for A.C.E. following small block cluster subsampling.

Layout: See Attachment B. Additions are:

- Segment Identifier
- Assignment Identifier
- Interview Flag

C. Sample Design File (Version 4 - Daily and Final)

Description: This file reflects the sampling through large block cluster subsampling. This file will be produced on a daily basis during large block cluster subsampling. The final version will be created when production is complete.

Level: Block Cluster

Scope: All block clusters selected in the initial A.C.E. sampling

File Layout: See Attachment C

D. Large Block Cluster Subsampling Segment File (Daily and Final)

Description: This file contains segment level information for all segments in all A.C.E. block clusters. This file will be produced on a daily basis during large block cluster subsampling. The final version will be created when production is complete.

Level: Block cluster segment

Scope: All block clusters that have more that one segment created during large block cluster subsampling.

File Layout: See section VI.G.3 above

VIII. VERIFICATION

The following information should be provided for verification:

A. Large Block Cluster Sampling Parameter File

Provide the sampling parameter file. The initial version of this file contains variables calculated from HU totals on the IL. Therefore, the initial version can be created and verified prior to forming and subsampling segments. Provide subsequent daily versions of this file during the large block cluster segment subsampling. See part A in section VII for the description of this file and Attachment A for the layout of this file.

B. Preliminary Enhanced List File

Make available the preliminary EL. Using this file, the Sample Design Team will verify HU totals, the assignment of subsampling status codes, segment creation, and workload assignments. Make this file available daily.

C. Updated Sample Design File and Large Block Segment File

Provide an updated version of the Design file and the large block segment file daily, and provide the final versions of these files at the end of the process. Using these files in conjunction with the sampling parameter file, DSSD will verify the implementation of the daily sampling.

IX. REFERENCES

1 DSSD Census 2000 Dress Rehearsal Memorandum Series A-9, "Census 2000 Dress Rehearsal ICM Sampling: Large Block Subsampling Specification," April 15, 1998.

2 DSSD Census 2000 Procedures and Operations Memorandum Series R-26, "Census 2000 Accuracy and Coverage Evaluation: Large Block Cluster Subsampling Parameter File Specification," March 8, 2000.

3 DSSD Census 2000 Procedures and Operations Memorandum Series R-3, "Accuracy and Coverage Evaluation (A.C.E.) Survey: Block Cluster Sample Selection Specification," March 29, 1999.

4 DSSD Census 2000 Procedures and Operations Memorandum Series R-, "Accuracy and Coverage Evaluation Survey: Reduction Specification," January 10, 2000, DRAFT.

5 DSSD Census 2000 Procedures and Operations Memorandum Series R-24, "Accuracy and Coverage Evaluation Survey: Small Block Cluster Subsampling," February 1, 2000.

6 DSSD Census 2000 Procedures and Operations Memorandum Series, Chapter S-HU-08, "Creation of the Census 2000 Accuracy and Coverage Evaluation (A.C.E.) Enhanced List for Person Phase Interviewing," June 21, 1999, DRAFT.

7 DSSD Census 2000 Procedures and Operations Memorandum Series, Chapter S-DT-01, "Accuracy and Coverage Evaluation: The Design Document," January 11, 2000.

8 DSSD Census 2000 Procedures and Operations Memorandum Series R-8, "Census 2000 Specifications for Block Cluster Formation-Reissue," May 3, 1999.

9 DSSD Census 2000 Procedures and Operations Memorandum Series R-4, "Accuracy and Coverage Evaluation (A.C.E.) Survey: Sample Summary File and Sample Design File Documentation," March 30, 1999.

cc: DSSD Census 2000 Procedures and Operations Memorandum Series Distribution List
A.C.E. Implementation Team/Statistical Design Team Leaders List
Sample Design Team

Layout of Sampling Parameter Files

The Initial Daily and Final Large Block Cluster Subsampling Parameter Files have the following file layout:

Variable Description	Name	Pos
State	ST	1-2
A.C.E. reduction stratum	ARST	4-5
Target housing unit sample size	T	7-14
Number of housing units in block clusters with 80 or more housing units on the independent list	NILHUL	16-21
Number of housing units in block clusters with 0-79 housing units on the independent list (except smalls with 0-9)	NILHUM	23-28
Number of housing units in all block clusters on the independent list	NILHUT	30-35
Number of housing units in small block clusters with 0-9 housing units on the independent list	NILHUS	37-42
Take-every for the segment subsampling	TELB	44-54
Number of segments in a block cluster	NSEG	56-57
Flag for formula used for calculating NSEG	FORMULA	59
Random Number between 0 and 1	RN	61-72
Random Start for the segment subsampling	RS	74-84
Current Daily Start	DS	86-96
Cumulative Cluster Count	CCC	98-100
Daily Start for Day 1	DS1	102-112
Daily Start for Day 2	DS2	114-124
.	.	.
.	.	.
.	.	.
Daily Start for Day 20[2]	DS20	

[2]The number of days for sampling may be over or under 20. If this is the case, appropriate modifications will be made.

Layout of the Preliminary Enhanced List

Layout Name : ENHANCED00.LAY
Description : 2000 ENHANCED LIST LAYOUT
Total Length : 360
Date Created : 11-22-1999

#	Field	Field description	length	Positions Beg	-	End	
1.	CNTRLNM	CONTROL NUMBER	24	1	-	24	CHAR
		1: 4 LCO					
		5:10 CLUSTER					
		11:12 SEGMENT					
		13:17 MAP SPOT NUMBER					
		18:21 WITHIN MSN ID					
		22:24 ZERO FILL					
2.	LCO	LOCAL CENSUS OFFICE	4	25	-	28	CHAR
		******************************					
		Index 1 CLUST thru WMSN					
		******************************					
3.	CLUST	CLUSTER NUMBER	6	29	-	34	CHAR
4.	MSN	ENHANCED IL MAP SPOT NUMBER	5	35	-	39	CHAR
5.	WMSN	WITHIN MAP SPOT NUMBER ID	4	40	-	43	CHAR
		******************************					
		Index 2 CID					
		******************************					
6.	CID	MAF ID	12	44	-	55	CHAR
7.	BLK	1998 BLOCK AND SUFFIX	6	56	-	61	CHAR
8.	URBNZ	URBANIZATION	30	62	-	91	CHAR
9.	HSNUM	HOUSE NUMBER (LJ/BF)	10	92	-	101	CHAR
10.	SNAME	STREET NAME (LJ/BF)	35	102	-	136	CHAR
11.	UNIT	UNIT DESIGNATION (LJ/BF)	15	137	-	151	CHAR
12.	RR	RURAL ROUTE/BOX # (LJ/BF)	25	152	-	176	CHAR
13.	POBX	PO BOX NUMBER (LJ/BF)	10	177	-	186	CHAR
14.	CITY	CITY/TOWN NAME	20	187	-	206	CHAR
15.	ZIP	ZIP CODE	5	207	-	211	CHAR
16.	ZIP4	ZIP + 4	4	212	-	215	CHAR
17.	STATE	FIPS STATE ABBREVIATION	2	216	-	217	CHAR
18.	FIPSCNTY	FIPS COUNTY CODE	3	218	-	220	CHAR
19.	FIPST	FIPS STATE CODE	2	221	-	222	CHAR
20.	PL	PHYSICAL LOCATION DESCRIPTION	50	223	-	272	CHAR
21.	PRKNM	TRAILER PARK NAME	30	273	-	302	CHAR
22.	HUFIN	MATCH CODE FROM HU MATCHING	2	303	-	304	CHAR
23.	HUFINID	ID FROM HOUSING UNIT MATCHING	12	305	-	316	CHAR
24.	TOA	TYPE OF BASIC ADDRESS	1	317	-	317	CHAR
		1 = ONE FAMILY HOUSE					
		2 = BSA WITH 2 OR MORE HUS					
		3 = MOBILE HOME NOT IN PARK					
		4 = MOBILE HOME IN PARK					
		5 = ONE FAMILY HOME IN SPECIAL PLACE					
		6 = BSA WITH 2 OR MORE HUS IN A SPECIAL PLACE					
		7 = OTHER					
25.	USTAT	UNIT STATUS	1	318	-	318	CHAR
		1 = OCCUPIED OR VACANT AND INTENDED FOR OCCUPANCY					
		2 = UNDER CONSTRUCTION					
		3 = FUTURE CONSTRUCTION					
		4 = UNFIT FOR HABITATION					
		5 = BOARDED UP					
		6 = STORAGE OF HOUSEHOLD					

```
Layout Name :  ENHANCED00.LAY                                    Page :    2
 Description :  2000 ENHANCED LIST LAYOUT
Total Length :    360
Date Created :  11-22-1999

                                                                   Positions
 #          Field           Field description              length  Beg  -   End
                                GOODS
                            7 = VACANT MOBILE HOME SITE
                            8 = OTHER
 26.  UR                 URBAN/RURAL                          1    319  -   319 CHAR
                            1 = URBAN
                          • 2 = RURAL
 27.  QAFLG              QA SAMPLE FLAG                       1    320  -   320 CHAR
                            0 = NOT IN QA SAMPLE
                            1 = IN QA SAMPLE
 28.  ESAMPFLG           E-SAMPLE ELIGIBILITY FLAG            1    321  -   321 CHAR
 29.  URFLAG             FLAG INDICATING THAT ADDRESS         1    322  -   322 CHAR
                           IS CONSIDERED TO BE URBAN OR
                           RURAL
                            0 = RURAL
                            1 = URBAN
 30.  MULTIFLAG          FLAG INDICATING THAT UNIT IS         1    323  -   323 CHAR
                           IN A MULTIUNIT OF LESS
                           THAN 20 UNITS
                            0 = MULTI <20 UNITS
                            1 = NONMULTI, OR MULTI >= 20
 31.  DSSDSEG            SEGMENT FOR LARGE BLOCK SUBSAM       2    324  -   325 CHAR
 32.  FLDSEG             SEGMENT FOR ASSIGNING WORK IN        2    326  -   327 CHAR
 33.  INTERVW            AFTER LARGE BLOCK SUBSAMP            1    328  -   328 CHAR
                            0 = OUT OF SAMPLE
                            1 = IN SAMPLE
 34.  JIC                JUST IN CASE SPACE                  10    329  -   338 CHAR
                         ******************************
                         THESE FIELDS ARE USED FOR LARG
                         BLOCK SUBSAMPLING.
                         ******************************
 35.  UNITCNT            NUMBER OF UNITS IN STRUCTURE         4    339  -   342 CHAR
 36.  TOTCASES           NUMBER OF CASES IN CLUSTER           6    343  -   348 CHAR
 37.  ICMCASES           NUMBER OF ICM CASES IN CLUSTER       6    349  -   354 CHAR
 38.  CENCASES           NUMBER OF CEN CASES IN CLUSTER       6    355  -   360 CHAR
```

Sample Design File

The Sample Design File contains one record per block cluster selected during the listing sample selection. If the block cluster falls out of sample during the second step of the listing sample, the A.C.E. reduction, small block cluster subsampling, or the A.C.E. reduction, the remaining variables will be left blank. The initial version of the file, which will be created following the initial block cluster selection, is called SDF.US1. For each subsequent update to the file, the version number will increase by one (i.e. SDF.US2, SDF.US3). The layout for the Sample Design File is as follows:

Variable Description	Name	Places	Source
Census Region	REGION	1	UN
Census Division	DIV	2	UN
State code	STATE	3-4	UN
County code	COUNTY	5-7	UN
Local census office	LCO	8-11	CS
Interim Tract (Pseudo Tract)	ITRACT	12-17	BC
Current Sample Indicator	CSI	19	UO
A.C.E. block cluster number	CLUST	21-25	CS
Check Digit	DIGIT	26	CS
Geography block cluster number	GCLUST	28-32	BC
List/Enumerate Indicator	LEIND	33	BC
Type of Enumeration Area Recode	TEACR	34	CS
Type of Enumeration Area group	TEAG	36	BC
Number of HUs used for sample design	NHU	37-41	BC
Number of MAF HUs	NHUM	43-47	BC
Number of 1990 HUs	NHU90	49-53	BC
Sampling Stratum 1 = Small 2 = Medium 3 = Large 4 = American Indian Reservation	SS	55	UN
American Indian Country Indicator 0 = No American Indian Country 1 = American Indian Reservation/trust land 2 = Tribal Jurisdiction Area/ Alaska Native Village Statistical Area/ Tribal Designated Statistical Area	AICIND	56	BC
Demographic/Tenure Group code	DTCODE	57-58	UN
Demographic/Tenure Group label	DTLABEL	59-60	UN
Estimated Urbanicity of block cluster 1 = Urban Area with population ≥250,000 2 = Other Urban Area 3 = Non-Urban Area	ECLUSURB	62	UN
Size Category 1=Small (0-2 hus) 2=Medium (3-79 hus) 3=Large (80+ hus)	SIZCAT	63	UN
Additional space		64-91	

Variable Description	Name	Places	Source
First step index number	INDEX1	92-99	CS
Listing sample selection indicator	BC1	101	CS
1 = Selected			
Random Start for listing sample selection	RS1	103-113	UN
Take-every for listing sample selection	TE1	115-125	UN
Second step listing sample selection indicator	BC2	127	CS
0 = Not Selected			
1 = Selected			
Random Start for the second step of the listing sampling	RS2	129-139	CS
Take-every for the second step of the listing sampling	TE2	141-151	CS
Unbiased weight after block cluster sampling	WEIGHTBC	153-164	CS
Additional space		165-175	
Preliminary Number of HUs on the Independent List	NHUILP	176-180	AR
Number of Housing Units On the January 2000 DMAF	NHUDMAF	182-186	AR
Demographic Code	DEMCODE	188	AR
1 = Minority			
2 = Non-Minority			
3 = Puerto-Rico			
Consistency Code	CONCODE	189	AR
1 = Low Inconsistent (IL significantly smaller than DMAF)			
2 = Consistent			
3 = High Inconsistent ((IL significantly larger than DMAF)			
A.C.E. Reduction Stratum	ARST	190-191	AR
A.C.E. Reduction Indicator	ACERED	193	AR
0 = Not Selected			
1 = Selected			
Random Start for A.C.E. Reduction	RSAR	195-205	AR
Take-every for A.C.E. Reduction	TEAR	207-217	AR
Unbiased weight after A.C.E. reduction	WEIGHTAR	219-230	AR
Collapsing Flag	COLFLAG	232	AR
A.C.E. Reduction Index Number	INDEXR	234-241	AR
Number of Housing Units On the December 1999 DMAF (Initial)	NHUDMAFI	243-247	AR
Additional space		248-300	
Number of HUs on the Independent List	NHUIL	301-305	SB
Small Block Cluster Subsampling Stratum	SBCSS	306-307	SB
Small Block Subsampling Indicator	SB	308	SB
0 = Not Selected			
1 = Selected			
Random Start for Small Block subsampling	RSSB	310-320	SB
Initial take-every for Small Block subsampling	ITESB	322-332	SB
Unbiased weight for A.C.E. cluster	WEIGHTC	334-345	SB
Larger of the DMAF and IL HU count	LARGERHU	347-351	SB
Final take-every for Small Block subsampling	FTESB	352-362	SB
Additional space		363-370	

Variable Description	Name	Places	Source
Relisted Block Cluster Flag 0 = Not Relisted, 1 = Relisted	RELIST	371	LB
Number of total hus in block cluster	NHUEL	373-377	LB
Number of A.C.E. hus in cluster	NHUELA	379-383	LB
Number of supplemental hus in cluster	NHUELN	385-389	LB
Large Block Cluster EL subsampling code 1 = NHUELI< 80 hus, 2 = NHUELI ≥ 80 hus	ELLBSUB	391	LB
Random Start for Large Block subsampling	RSLB	393-403	LB
Take-every for Large Block subsampling	TELB	405-415	LB
Number of segments in block cluster	NSEG	417-418	LB
Number of segments selected in block cluster	NSEGSAM	420-421	LB
Day of Arrival	DAY	423-424	LB
Final Cluster Order Number	CON	431-434	LB
Number of total hus for interview in block cluster	NINT	436-440	LB
Unbiased weight for P-sample HUs	WEIGHTP	442-453	LB
Number of Assignments in block cluster	NA	455-456	LB
Final Sampling Strata	FSS	458-464	LB
Additional space		465-490	

Source Codes

AR: A.C.E. Reduction
BC: Block Clustering
CS: Block Cluster Sampling
LB: Large Block Subsampling
SB: Small Block Subsampling
UN: Universe File Creation
UO: Updated for each operation

Large Block Cluster Subsampling Example

This hypothetical example demonstrates the phases that a block cluster goes through during the large block cluster subsampling process.

1. Calculate the Sampling Parameters for Each A.C.E. Reduction Stratum and State (see reference 2)

   Sampling parameter calculations occur prior to the arrival of block clusters. This information is based on results from the IL. Let's say for a particular state and A.C.E. reduction stratum, the target number of HUs, T, the number of HUs in block clusters with more than 80 A.C.E. HUs, NILHUL, and the number of HUs in block clusters with 0-79 A.C.E. HUs (except small block clusters with less than 10 IL HUs), NILHUM, are:

   T = 2050
   NILHUL = 1295
   NILHUM = 1173

   The sampling parameters calculated from this information are the take-every, TELB, the number of segments per block cluster, NSEG, and the random start, RS.

$$\text{TELB} = \frac{\text{NILHUL}}{\text{T} - \text{NILHUM}} = \frac{1295}{2050 - 1173} = 1.477000$$

$$\text{NSEG} = \frac{1}{1 - \frac{1}{\text{TELB}}} = \frac{1}{1 - \frac{1}{1.477000}}$$ = 3.096000 (Rounded up to the next integer) = 4

   FORMULA = 2, since formula 2 was used.

A random number is selected, RN = 0.179317, and the random start is calculated.

$$\text{RS} = \text{RN} \times 1.477 = 0.179317 \times 1.477000 = 0.265000$$

The remaining phases in large block cluster subsampling rely on information obtained from the HUs on the preliminary EL. Suppose two block clusters from the same A.C.E. reduction stratum within a state arrive on day one. These block clusters go through large block cluster processing as follows:

2. Create Block Cluster Housing Unit Counts and Subsampling Status Codes

To determine whether each block cluster needs to be segmented and subsampled or the entire cluster remains in sample, HU counts are calculated. The two types of HUs are A.C.E. HUs, NHUELA, and supplemental HUs, NHUELN. The combination of these two types is the total HUs on the preliminary EL, NHUEL. The subsampling codes, ELLBSUB, SEGSUB, SEGSAM, and SEGID are assigned to the block clusters that do not need to be segmented and subsampled.

Block Cluster	NHUEL	NHUELA	NHUELN	ELLBSUB	SEGSUB	SEGSAM	SEGID	In AIR?	Next Phase
1	247	197	50	1	1	1	AA	yes	Send to 6
2	83	82	1	2	TBD	TBD	TBD	no	Send to 3

3. Create Segments

For those non-AIR block clusters with 80 or more A.C.E. HUs (block cluster 2), segments need to be formed. The most that the segments within a block cluster will differ is one A.C.E. HU. A segment identifier is also assigned to distinguish between segments in each block cluster.

$$AVGSEG = trunc(\frac{NHUELA}{NSEG})$$

$$R = NHUELA - NSEG \times AVGSEG$$

The supplemental HUs are assigned to the same segment as the preceding A.C.E. HU based on map spot number. For this example, suppose the one supplemental HU is assigned to the segment CA.

Block Cluster	Avg.	AVGSEG	R	Segment AA A.C.E. HUs (Supplemental HUs)	Segment BA A.C.E. HUs (Supplemental HUs)	Segment CA A.C.E. HUs (Supplemental HUs)	Segment DA A.C.E. HUs (Supplemental HUs)
2	20.5	20	2	21 (0)	21 (0)	20 (1)	20 (0)

4. Create Segment Level Variables and Codes

HU counts were calculated previously for block clusters. In this section, HUs are counted similarly for each segment. A.C.E. HUs, supplemental HUs, and total HUs are counted.

Block Cluster	SEGID	NHUELS	NHUELAS	NHUELNS
2	AA	21	21	0
	BA	21	21	0
	CA	21	20	1
	DA	20	20	0

5. Select a Systematic Sample of Segments for Each A.C.E. Reduction Stratum and State

After the non-AIR block clusters with 80 or more HUs have been segmented, a systematic sample of segments is selected using the parameters calculated in step 1. The sampling is across all segments in all block clusters within the same A.C.E. reduction stratum, state, and day. Since the selection is done over several days, the systematic sample needs to be carried over from day to day. A daily segment order number (DSON) is assigned to implement the sampling on a daily basis. The cluster order number (CON) over all days is also assigned. Information to be carried over are the daily end value (DE), which is the following day's start value (DS), and the cumulative cluster count (CCC), which is a cumulative count of block clusters processed.

So for day 1:

TE = 1.477000 and DS = 0.265000

L1 = DS = 0.26500	→ take segment 1
L2 = 1.742000	→ take segment 2
L3 = 3.219000	→ take segment 4

Day	CON	SEGID	DSON	SEGSUB
1	1	AA	1	1
		BA	2	1
		CA	3	0
		DA	4	1

N = 4 and n = 3

DE = 0.265000 + 1.477000×3 - 4 = 0.696000 →start day 2 with a DS = 0.696000

CCC = 1 →start day 2 with a CON = CCC + 1 = 2

6. Identify Housing Units for A.C.E. Interview

If an A.C.E. HU has an INTERVW of 1, then that HU is sent for interview. If an A.C.E. HU has an INTERVW of 0, then that HU is not sent for interview. Supplemental HUs have an INTERVW of 8 or 9 and none of them will be sent for interview.

Block Cluster	Segment	INTERVW	Result
1	AA	1 for A.C.E. HUs 9 for supp. HUs	Interview 197 A.C.E. HUs
2	AA	1 for A.C.E. HUs	Interview 21 A.C.E. HUs
	BA	1 for A.C.E. HUs	Interview 21 A.C.E. HUs
	CA	0 for A.C.E. HUs 8 for supp. HU	Interview 0 A.C.E. HUs
	DA	1 for A.C.E. HUs	Interview 20 A.C.E. HUs

7. Create Interviewer Workload Assignments

After determining the number of HUs to be sent for interview in a block cluster, the next step is to determine the number of assignments. A block cluster with 80 or more HUs to be sent for interview is divided into two or more assignments of about 40 to 50 HUs per assignment. Block clusters with 0-79 HUs to be sent for interview are left as one assignment. The number of HUs for interview is 197 in block cluster 1 and 62 in block cluster 2, thus only block cluster 1 needs to be split into assignments.

$$NA = \frac{NINT}{40} = \frac{197}{40} = 4.9 \text{ (Rounded down)} = 4 \text{ Assignments}$$

$$AVGHUA = \mathrm{trunc}(\frac{NINT}{NA}) = \mathrm{trunc}(\frac{197}{4}) = 49$$

$$RINT = NINT - NA \times AVGHUA = 197 - 4 \times 49 = 1$$

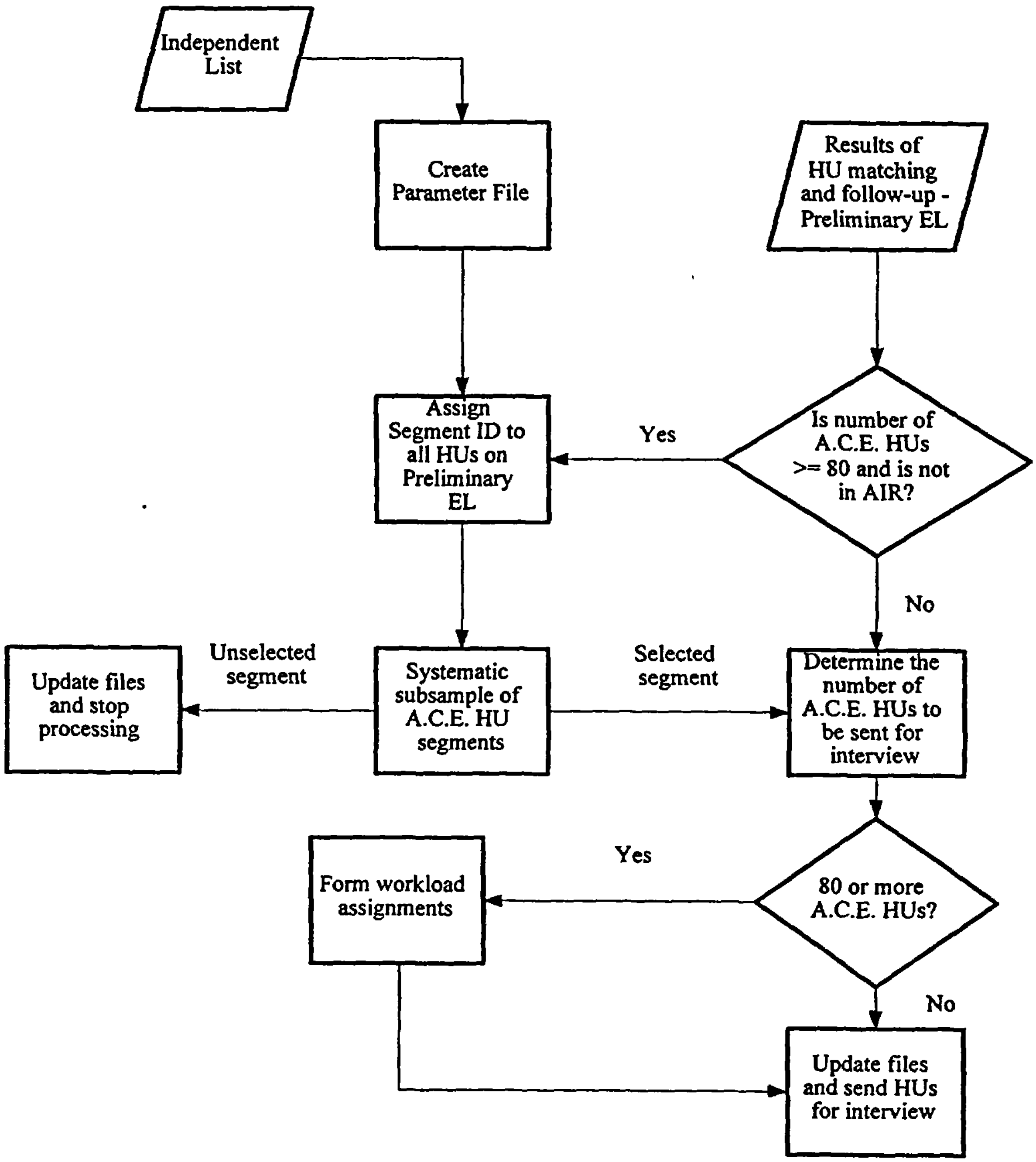
Give the first assignment one additional HU than the last three assignments.

For the first assignment:
ASIZE = 49 + 1 = 50 interviews
For the next three assignments:
ASIZE = 49 interviews

To distinguish between assignments within a block cluster ASSIGNID is assigned. For block cluster 2, ASSIGNID is set to AA since it did not require multiple workload assignments.

Block Cluster	Assignment	ASSIGNID	Interviews
1	1	AA	50
	2	AB	49
	3	AC	49
	4	AD	49
2	1	AA	62

Overview of the Large Block Cluster Subsampling Process



UNITED STATES DEPARTMENT OF COMMERCE
Bureau of the Census
Washington, DC 20233-0001

MASTER FILE

March 21, 2000

DSSD CENSUS 2000 PROCEDURES AND OPERATIONS MEMORANDUM SERIES R-29

MEMORANDUM FOR Dennis Stoudt
Assistant Division Chief, Processing Systems
Decennial Systems and Contracts Management Office

From: Donna Kostanich DK
Assistant Division Chief, Sampling and Estimation
Decennial Statistical Studies Division

Prepared by: Peter P. Davis PPD
Sample Design Team
Decennial Statistical Studies Division

Subject: Accuracy and Coverage Evaluation Survey: Reduction Specification

I INTRODUCTION

This memorandum describes the selection of the Accuracy and Coverage Evaluation (A.C.E.) reduction sample. The A.C.E. reduction is the sampling operation that links the original Integrated Coverage Measurement (ICM) Survey sampling plan to the A.C.E. sampling plan. The A.C.E. reduction is the first of several operations that will reduce the number of housing units (HUs) from the nearly two million HUs in independent listing to the approximately 300,000 HUs that will be sent for interview. Block clusters were selected for A.C.E. independent listing under the previously planned 750,000 HU ICM design. (See References 1, 2, 3, 4, and 5.) Since not all of the listed block clusters are required for A.C.E., the reduction will subsample those clusters, with the selected clusters continuing on in A.C.E. operations. The A.C.E. reduction is a subsample of medium and large block clusters in the 50 states and the District of Columbia. Small block clusters and block clusters on American Indian Reservations (AIR) and in Puerto Rico are not eligible for this reduction. The two remaining sampling operations are small block cluster subsampling and large block cluster subsampling.

In small block cluster subsampling, the 5,000 small block clusters selected in the A.C.E. listing sample will be subsampled following the keying of the independent listing books. As a general rule, if a cluster is larger than expected, then the cluster will be retained at a higher rate. This stage of sampling should not have a significant impact on the interview sample size but is critical for controlling the size of the cluster weights and for maximizing field interviewing efficiency.

The final stage of selection is large block cluster subsampling. When a cluster has more than 80 listed HUs, segments of adjacent HUs are formed and a subsample of segments is selected from the cluster. This produces the final A.C.E. interview sample.

The sections of this specification are ordered as follows:

- Section II: Overview
- Section III: Definitions
- Section IV: Assumptions
- Section V: Input Files
- Section VI: Output Files
- Section VII: Reduction Process
- Section VIII: References

Note that the A.C.E. reduction operation is complete. This specification reflects the reduction as it was actually implemented. The results of the reduction are documented in reference 6, and a contingency plan that addresses the delay of an updated census address list is documented in reference 7. This final specification includes changes to the original working draft that were required due to issues that arose in testing or production. The final specification retains the flexible approach to the reduction that was presented in the working draft. This approach was required since system development had to begin before a final sample design could be determined. Therefore, the reduction parameters are treated as variables in this specification even though in the final sample design they may have ended up as constants. For example, all medium stratum jumper clusters, those that moved from the medium sampling stratum to the large, were retained in the reduction sample even though their probability of selection is treated as a variable in this specification. In general, this final specification is nearly identical to the original working draft.

Any comments or questions should be directed to Pete Davis (301-457-8322), Jim Farber (301-457-4282), or Debbie Fenstermaker (301-457-4195) of the Decennial Statistical Studies Division (DSSD).

II PROCESS OVERVIEW

This overview will detail the steps of the A.C.E. reduction process. The steps listed below correspond to the steps in section VII, which contains programming instructions and is significantly less descriptive than this overview.

A. Read in the Sample Design File

The Sample Design File contains data for each block cluster selected in the first step of the listing sample. This file is the primary input file for the A.C.E. reduction since the sampling frame for the reduction is contained on the Sample Design File. Note that there are more clusters on the Sample Design File than were actually in the listing sample. This is due to the second step of listing sample selection, which was used to control expected listing workloads. The second step was needed only in Indiana and Missouri. The Current Sample Indicator (CSI) is used to screen out these clusters, which are not eligible for the A.C.E. reduction.

B. Assign Cluster Codes

Block clusters on the Sample Design File will be assigned two cluster codes for the A.C.E. reduction: Demographic strata codes and Consistency strata codes.

Demographic strata codes are based on the original Demographic/Tenure code assigned for listing sample selection (See Reference 4). The Demographic/Tenure code represents a classification of block clusters according to the approximate distribution of race/Hispanic origin and tenure, and was used as a sort variable in the selection of the listing sample. For the A.C.E. reduction, the 14 Demographic/Tenure codes will be combined into three Demographic strata, which are more than a sort variable in the reduction since sampling rates may vary across these strata. The three Demographic strata are:

- Minority: a block cluster with any minority (non-Other and non-Puerto Rico) Demographic/Tenure code
- Non-minority: a block cluster with any Other Demographic/Tenure code
- Puerto Rico: a block cluster with a Puerto Rico Demographic/Tenure code

Consistency strata codes are based on cluster HU count differences. An estimated HU count was created for listing sample selection based on the most recent Master Address File (MAF) HU counts available at the time. For the A.C.E. reduction, two updated cluster HU counts will be used: the preliminary independent listing (PIL) HU count and the Decennial Master Address File (DMAF) HU count. The PIL HU count is a preliminary HU count clerically tallied from the Independent Listing Book for each cluster in the listing sample. The DMAF HU count used in the A.C.E. reduction is taken from the January, 2000 version of the DMAF, which includes September and November Delivery Sequence File (DSF) updates. The PIL HU count and the DMAF HU count are compared and clusters are placed into consistency strata based on the relationship of those HU counts. Large differences between these counts indicate that coverage problems may occur and thus the weights for such clusters should be controlled to avoid serious variance effects.

Clusters will be placed into three consistency strata:

- Low Inconsistent: a block cluster where the PIL HU count is more than 25 percent lower than the DMAF HU count. Low Inconsistent clusters may have a large percentage of erroneous enumerations in the census.
- Consistent: a block cluster where the absolute difference between the PIL HU count and the DMAF HU count is not more than 25 percent.
- High Inconsistent: a block cluster where the PIL HU count is more than 25 percent higher than the DMAF count. High Inconsistent clusters may have a large percentage of omissions in the census.

The percent differences to use as cutoffs to define the consistency strata are specified as the parameters X_L and X_H on the Reduction Parameter File.

For List/Enumerate clusters, the DMAF HU count will not be known at the time of the reduction. Thus, all such clusters will be classified as High Inconsistent.

C. Stratify Clusters

Clusters will be placed into A.C.E. reduction strata based on the Demographic and Consistency codes created in step B above and based on the collapsing pattern for a given state. Collapsing is required since some states do not have the sample to support a full crossing of Demographic and Consistency strata. The collapsing pattern will be predefined for each state on the Reduction Parameter File.

Medium stratum jumpers, those clusters that had been in the medium sampling stratum for the listing sample but now have 80 or more PIL HUs, have their own A.C.E. reduction stratum. Medium clusters were sampled at lower rates than large clusters in the listing sample since large clusters eventually undergo large block cluster subsampling, an operation that increases their weights. Medium stratum jumpers will also go through large block cluster subsampling, meaning the already high weights of these clusters will become even larger. By taking all or most of the medium stratum jumpers in the A.C.E. reduction, their weights will be controlled and these clusters will not introduce the significant weight variation they otherwise would have. Similarly, small block clusters that have 80 or more PIL HUs have their own A.C.E. reduction stratum to facilitate controlling their weights during large block cluster subsampling.

AIR and Puerto Rico block clusters will also be placed in their own A.C.E. reduction strata. Small block clusters that do not have 80 or more PIL HUs will be assigned an A.C.E. reduction stratum code similar to medium and large clusters in order to compute reduction stratum target interview sample sizes for large block cluster subsampling.

For the complete set of reduction strata, see Attachment A.

D. Identify Eligible Clusters

Only certain block clusters are eligible for the A.C.E. reduction. The ineligible clusters will be removed from the process at this point. Information about these clusters, such as A.C.E. reduction stratum, will be saved to the Sample Design File for use in later A.C.E. sampling operations. The ineligible clusters are:

- Small block clusters
- AIR block clusters
- Puerto Rico block clusters

All other clusters will continue in the A.C.E. reduction process.

E. Calculation of Sampling Parameters

Prior to reduction, the Sample Design Team determined differential sampling rates for each reduction stratum relative to a baseline reduction stratum for each state. The differential sampling factors differ from reduction strata to reduction strata and from state to state depending on conditions such as the available sample in a state. The differential sampling factor for each stratum and state is provided on the Reduction Parameter File. The sample allocated to each A.C.E. reduction stratum accounts for the PIL HU count and the amount of differential sampling targeted for that stratum. If no differential sampling is desired in a state, the sample allocation is proportional to the PIL HU count in each stratum.

In addition to the different sampling rates among the A.C.E. reduction strata, medium and large clusters may be sampled at different rates. These different rates will occur in states where a listing adjustment or a second step of sampling was needed during the listing sample selection. The listing adjustment and the second step of sampling were two procedures built into the listing sample selection to control the expected number of HUs to list. These two procedures affected only large block clusters. Therefore, to restore the proportionality of the weights between the medium and large clusters, the take-everys for the large clusters will be adjusted to account for the listing adjustment and the second-step take-every.

Medium stratum jumpers have a predetermined take-every that is provided on the Reduction Parameter File. HUs in medium stratum jumper clusters are excluded from the take-every calculations for the other A.C.E. reduction strata. It is likely that all medium stratum jumpers will be retained in the sample unless an overly large number of clusters are medium stratum jumpers. In this case, a subsample of medium stratum jumpers will be taken in the same manner used for subsampling all other eligible clusters.

In the reduction process, two take-everys are calculated: an initial take-every and a final take-every. The final take-every is calculated to facilitate variance estimation by providing an integer block cluster sample size for each reduction stratum and state.

F. Select a Subsample of Block Clusters

Following the calculation of the sampling parameters, a systematic sample of clusters is selected for the medium clusters and large clusters separately within each A.C.E. reduction stratum and state. Selected block clusters remain in sample and continue to the next sampling operation.

G. Update Files

After the sample selection, output files are updated or created. The first is the Sample Design File, which records results from each of the sampling operations for each block cluster. The second file is the housing unit sample size file, which contains initial target interview sample sizes for each reduction stratum.

III DEFINITIONS

A. States

All 50 states, the District of Columbia, and Puerto Rico are "states" in the A.C.E. Reduction.

B. Sampling Stratum

Block clusters are classified into four sampling strata for the listing sample selection based on an early census count of HUs used for clustering. These categories are as follows:

1. Small block clusters: 0 - 2 HUs.
2. Medium block clusters: 3 - 79 HUs and not on an AIR.
3. Large block clusters: 80 or more HUs and not on an AIR.
4. American Indian Reservation block clusters: 3 or more HUs and on an AIR.

The sampling strata above are the original sampling strata as located on the Sample Design File in the variable SS, Sampling Strata.

C. Preliminary Independent Listing HU Count

The independent listing HU count used in the A.C.E. reduction is preliminary because it is simply a clerical tally of HUs from the independent listing books. These A.C.E. HU counts are obtained from the Cluster Count File, a file provided to the Decennial Systems and Contracts Management Office (DSCMO) from the Technologies Management Office (TMO). See Reference 8 for the specifications on independent listing file transfers.

D. Decennial Master Address File HU Counts

Census HU counts are obtained from the DMAF. The variables GQFLG (Group Quarters HU Flag) and SMAFID (Surviving MAFID) from the DMAF represent the characteristics of an address used in identifying DMAF HUs for the purposes of the A.C.E. reduction. The GQFLG code distinguishes Census HUs from group quarters. The SMAFID code identifies duplicate HUs on the DMAF. Possible values for GQFLG and SMAFID are as follows:

GQFLG:
0 = Housing Unit
1 = Special Place
2 = Group Quarters
3 = GQ Embedded Housing Unit

SMAFID:
0 = address is not a duplicate. (A non-zero SMAFID implies the address is a duplicate.)

The GQFLG and SMAFID codes that will be recognized as valid DMAF HUs for A.C.E. reduction are as follows:

GQFLG = 0 and SMAFID = 0, or
GQFLG = 3 and SMAFID = 0.

E. Consistency Strata

Consistency strata are groups of clusters formed on the basis of the percent difference between the PIL HU count and the DMAF HU count. A cluster is allocated to a consistency stratum by the definitions in Table 1:

Table 1. Consistency Strata Definitions

Consistency Stratum	If	Then
	Criteria	Consistency Code
Low Inconsistent	PIL < $X_L \times$ DMAF	1
Consistent	$X_L \times$ DMAF $\leq$ PIL $\leq X_H \times$ DMAF	2
High Inconsistent	PIL > $X_H \times$ DMAF	3

The variables X_H and X_L are "Inconsistency Cutoffs." X_H is the high inconsistency cutoff ($X_H = 1.25$ meaning 25 percent higher than the PIL) and X_L is the low inconsistency cutoff ($X_L = 0.75$ meaning 25 percent lower than the PIL). These cutoffs are specified on the Reduction Parameter File. At the time of A.C.E. reduction, List/Enumerate clusters do not have a DMAF HU count. Hence, all List/Enumerate clusters are in the High Inconsistent Stratum.

F. Demographic Strata

Demographic strata are groups of clusters formed on the basis of the estimated 1990 racial and Hispanic ethnicity distribution. Using the Demographic/Tenure Group Code assigned during the listing sample selection, labeled DTCODE on the Sample Design File, a cluster is allocated to a demographic stratum based on the definition in Table 2:

Table 2. Demographic Strata Definitions

Demographic Stratum	IF	THEN
	Criteria	Demog. Strata Code
Minority	DTCODE = 1, 2, 3, 4, 5, 6, 7, 8, 9, or 10	1
Non-minority	DTCODE = 11 or 12	2
Puerto Rico	DTCODE = 13 or 14	3

G. A.C.E. Reduction Strata

Block clusters are classified into one of nineteen (19) A.C.E. reduction strata based on original sampling stratum code, demographic stratum code, consistency stratum code, and a collapsing flag. See Attachment A for a complete list of all nineteen A.C.E. reduction strata.

H. Stratum Jumpers

Between the time of selecting the listing sample and the A.C.E. reduction, it is possible for clusters to change from their original sampling stratum. Such clusters are referred to as Stratum Jumpers. The only stratum jumpers of interest at this point, however, are the medium stratum to large stratum jumpers because these clusters require special attention in the reduction to control their weights. Medium stratum jumpers are clusters that are in the medium sampling stratum on the Sample Design File but have a PIL HU count of 80 or more. Small stratum jumpers, those clusters in the small sampling stratum at the time of listing but with a PIL HU count of 80 or more, are also identified in A.C.E reduction but are dealt with during small block cluster subsampling and large block cluster subsampling.

I. A.C.E. Reduction Parameters

The DSSD will provide certain parameters needed for the A.C.E. reduction on the Reduction Parameter File (See Attachment B for a layout).

1. Differential Sampling Factors

The differential sampling factors are the sampling rates relative to the baseline A.C.E. reduction stratum. The baseline A.C.E. reduction stratum is the stratum with the lowest sampling rate. In every state, the Consistent stratum is the baseline. The differential sampling factors indicate the degree to which the Minority, Inconsistent Low, and Inconsistent High reduction strata are differentially sampled relative to the Non-Minority Consistent reduction stratum. For example, a differential sampling factor of two for the Minority reduction stratum means that the probability of selection for a cluster in the Minority stratum is twice that of a cluster in the Consistent stratum.

2. Inconsistency Cutoffs

The Inconsistency Cutoffs are the critical values (in terms of percentages) that define significant differences between the PIL and DMAF HU counts. For the purposes of the reduction, a significant difference is more than 25 percent. Significant differences can occur on both the high end and the low end, hence the parameters X_H and X_L.

3. Listing Adjustment

The listing adjustment is a factor that was applied to increase the first-step take-every of the large sampling stratum during listing sample selection to reduce the expected listing workload in some states. The A.C.E. reduction will compensate for the decrease of large clusters in the listing sample due to the listing adjustment.

4. Second-step Take-every

During listing sample selection, a second step of sampling was required in some states in the large sampling stratum to reduce the expected listing workload. The A.C.E. reduction will compensate for the second-step sampling of large clusters.

5. Collapsing Flag

The collapsing flag is a flag on the Reduction Parameter File that indicates the prespecified collapsing pattern that will be used to form the A.C.E. reduction strata in each state. The collapsing flag represents a strategy to assign an A.C.E. reduction stratum to a block cluster based on its original sampling stratum and demographic/consistency characteristics. An example of the use of the collapsing flag is illustrated on page 17.

6. Medium and Large Block Cluster Weights after Listing Sample Selection

Large clusters were sampled at a higher rate than medium clusters during the listing sample selection, creating different weights for medium and large clusters. These weights will be required to calculate reduction sampling parameters.

IV ASSUMPTIONS

A. The A.C.E. listing sample of block clusters was selected according to the previously planned ICM 750,000 HU design. There are 29,695 block clusters in the listing sample including 559 in Puerto Rico.

B. Independent listing HU counts are preliminary. This is due to time constraints. Any reference to the Independent List at this stage in the A.C.E. sample survey will be referred to as the Preliminary Independent List (PIL). Post-A.C.E. reduction processes such as small block cluster subsampling and large block cluster subsampling will have a "Keyed and Valid" independent listing of HUs. "Keyed and Valid" implies these HU counts will have undergone a complete quality control.

C. The final A.C.E. sample size is approximately 300,000 HUs. This sample size is for interview after the A.C.E. reduction, small block cluster subsampling, and large block cluster subsampling.

D. Only medium and large block clusters are sampled in the A.C.E. Reduction. All small block clusters, American Indian Reservation block clusters, and Puerto Rico block clusters remain in the sample.

E. A.C.E. reduction stratum codes are stored in two-digit fields and are zero-filled as needed.

F. Decimal numbers are rounded to six (6) digits unless otherwise noted at the time of creation using the standard rounding procedure. Standard rounding in this specification means that a number with a seventh-decimal value of five or higher is rounded up in the sixth decimal. Otherwise, the sixth decimal value is unchanged.

V INPUT FILES

The following files will be used in this process.

A. Reduction Parameter File

This file contains the parameters for each state that are needed for the A.C.E. reduction. This file will be provided by the DSSD. There is one record for each A.C.E. reduction stratum within each state. See Attachment B for the file layout and an example of the data in the Reduction Parameter File.

B. Decennial Master Address File (DMAF)

This file contains address information for each HU in Census 2000. The DMAF is formed from an extract of the MAF along with updates from the United States Postal Service DSF and from census operations such as Local Update of Census Addresses. See Reference 9 for more information on the DMAF.

C. Cluster Count File

The Cluster Count File contains one record for each of the 29,695 block clusters in the A.C.E. listing sample. The TMO will transmit a file to the DSCMO at the end of the independent listing operation. Each record will contain a preliminary count of the HUs listed in each cluster during independent listing.

Attachment C contains the layout of this file. For further details on this file, see Reference 8.

D. Sample Design File

The Sample Design File contains one record per block cluster chosen during the first step of listing sample selection. This file tracks the path that each block cluster travels during the A.C.E. sampling procedures. The Sample Design File contains categorical variables corresponding to each procedure as well as parameters and HU totals. If the block cluster fell out of sample at some point, the remaining variables are left blank. The variable CSI is used to indicate which block clusters are in sample; clusters with a CSI of one are in the sample. The initial version of the file, which was created following the listing sample selection and is the input for the A.C.E. reduction, is called SDF.US1. There are 29,717 records on the Sample Design File. Attachment D contains a file layout for the Sample Design File.

VI OUTPUT FILES

A. Housing Unit Sample Size File

The HU Sample Size File contains three variables for large block cluster subsampling: State, A.C.E. Reduction Stratum, and Target Number of HUs.

Variable Description	Name	Location
State code	STATE	1-2
A.C.E. Reduction Stratum (zero-filled)	ARST	4-5
Target Number of HUs to interview in A.C.E. reduction stratum	T	7-14

B. Sample Design File

Updates will be made to the Sample Design File based on the results of the A.C.E. reduction. After all states have been verified, the new version of the Sample Design File will be called SDF.US2.

C. Reduction Parameter File

Updates will be made to the Reduction Parameter File during the A.C.E. reduction process so that the Sample Design Team in the DSSD may check the parameters for statistical validity.

VII REDUCTION PROCESS

Process each state as follows:

A. Read in Sample Design File

The layout for the Sample Design File is located in Attachment D. In the current state, read in the following fields for each cluster with CSI = 1:

Variable Description	Name	Location
State code	STATE	3-4
Current Sample Indicator	CSI	19
A.C.E. block cluster number and check digit	CLUST	21-26
List/Enumerate Indicator	LEIND	33
Sampling Stratum	SS	55
Demographic/Tenure group code	DTCODE	57-58
First-step index number	INDEX1	92-99
Unbiased weight after listing sample	WEIGHTBC	153-164

B. Assign Cluster Codes

1. Demographic Strata Codes

Assign the demographic stratum code created below to each cluster with CSI = 1 on the Sample Design File. Using DTCODE from the Sample Design File, assign DEMCODE to each cluster using the rules in Table 3:

Table 3. Demographic Stratum Code Assignment Rules

Demographic Stratum	IF	THEN
	Criteria	DEMCODE
Minority	DTCODE = 1, 2, 3, 4, 5, 6, 7, 8, 9, or 10	1
Non-minority	DTCODE = 11 or 12	2
Puerto Rico	DTCODE = 13 or 14	3

2. Consistency Strata Codes

Assign a consistency stratum code to each cluster with CSI = 1 using the following steps:

a. Obtain the following information for each cluster:

- the DMAF HU count from the January, 2000 DMAF as defined in section V.B, NHUDMAF
- the DMAF HU count from the December, 1999 DMAF, NHUDMAFI (See Reference 7 for more information on the two DMAF versions.)
- the PIL HU count from the Cluster Count File, NHUILP
- the Inconsistency Cutoffs for the current state from the Reduction Parameter File, X_H and X_L.

b. For each List/Enumerate (L/E) cluster on the Sample Design File, assign a consistency stratum code, CONCODE, equal to "3", High Inconsistent. The L/E clusters have LEIND = 1 on the Sample Design File.

c. For each non-L/E cluster, assign a Consistency Stratum Code, CONCODE, using the PIL HU count, the DMAF HU count, and the Inconsistency Cutoffs according to the rules in Table 4:

Table 4. Consistency Stratum Code Assignment Rules

Consistency Stratum	If	Then
	Criteria	CONCODE
Low Inconsistent	PIL < $X_L \times$ DMAF	1
Consistent	$X_L \times$ DMAF $\leq$ PIL $\leq$ $X_H \times$ DMAF	2
High Inconsistent	PIL > $X_H \times$ DMAF	3

C. Stratify Clusters

Assign each cluster with CSI = 1 on the Sample Design File to an A.C.E. reduction stratum using the following hierarchical rules:

1. Assign all AIR clusters an A.C.E. reduction stratum of "17". These are clusters in the AIR sampling stratum (SS = 4) on the Sample Design File.

2. Assign all medium stratum jumpers an A.C.E. reduction stratum of "16". Medium stratum jumpers are clusters that were originally in the medium sampling stratum (SS = 2) for the listing sample but have a PIL HU count of 80 or more.

3. Assign all small stratum jumpers an A.C.E. reduction stratum of "19". Small stratum jumpers are clusters that were originally in the small sampling stratum (SS = 1) for the listing sample but have a PIL HU count of 80 or more.

4. Assign all Puerto Rico clusters that are not medium or small stratum jumpers an A.C.E. reduction stratum of "18". These are clusters having a FIPS state code (STATE) of "72".

5. Obtain the Collapsing Flag for the current state from the Reduction Parameter File.

6. Assign A.C.E. reduction strata codes to all other clusters using Table 5 below:

Table 5. A.C.E. Reduction Strata Assignments

DEMCODE	CONCODE	Collapsing Flag								
		01	02	03	04	05	06	07	08	09
1	1	01	01	01	07	10	11	11	13	14
1	2	01	01	01	08	08	12	12	13	12
1	3	01	01	01	09	10	11	11	13	15
2	1	02	05	06	07	10	02	05	13	02
2	2	03	03	06	08	08	03	03	13	03
2	3	04	05	06	09	10	04	05	13	04

The strategy of Table 5 is, given a collapsing flag for a state, work down the column of the collapsing flag until the appropriate DEMCODE and CONCODE are reached and assign the cluster to the A.C.E. reduction stratum in the corresponding cell of the table.

As an illustration, given a cluster with a DEMCODE of "2" implying a non-minority cluster, a CONCODE of "3" meaning a High Inconsistent cluster, and a collapsing flag of "05" for the state, then the A.C.E. Reduction Stratum is "10." Using Attachment A, an A.C.E. reduction stratum of "10" is called "Inconsistent". Therefore, this A.C.E. reduction stratum was collapsed over demographics (minority and non-minority) and over inconsistency (high and low) to include all "Inconsistent" clusters.

D. Identify Eligible Clusters

At this point, clusters will be identified as eligible or ineligible for further processing in the A.C.E. reduction. Small block clusters, AIR block clusters, and Puerto Rico block clusters are ineligible, while all other clusters are eligible.

1. For all block clusters with SS = 1, SS = 4, or STATE = 72, update the following variables on the Sample Design File. Set TEAR = 1.000000, RSAR = 1.000000, ACERED = 1, WEIGHTAR = WEIGHTBC,

INDEXR = "   ", and COLFLAG = the collapsing flag for that state from the Reduction Parameter File.

Variable Description	Name	Location
Preliminary Number of HUs on the Independent List	NHUILP	176-180
Number of HUs on the January 2000 DMAF	NHUDMAF	182-186
Demographic Code	DEMCODE	188
Consistency Code	CONCODE	189
A.C.E. Reduction Stratum (zero-filled)	ARST	190-191
A.C.E. Reduction Indicator	ACERED	193
Random Start for A.C.E. Reduction	RSAR	195-205
Take-every for A.C.E. Reduction	TEAR	207-217
Unbiased weight after A.C.E. reduction	WEIGHTAR	219-230
Collapsing Flag	COLFLAG	232
A.C.E. Reduction Index Number	INDEXR	234-241
Number of HUs on the December 1999 DMAF (Initial)	NHUDMAFI	243-247

For ineligible clusters, the A.C.E. reduction process terminates at this point.

2. For all other block clusters, continue to step E below.

E. Calculation of Sampling Parameters

1. Obtain the Variables for Calculating Take-Everys and Expected Sample Sizes

a. Obtain the following information for the current state from the A.C.E. Reduction Parameter File:

i. Differential sampling factors for each A.C.E. reduction stratum, K
ii. Listing Adjustment, LISTADJ
iii. Second step take-every, TE2
iv. Medium stratum jumper take-every, TESJ
v. Medium cluster weight after the listing sample selection, WTM
vi. Large cluster weight after the listing sample selection, WTL
vii. Target number of housing units for interview in the current state, THU. Note: THU does not include HUs in stratum jumper clusters.

b. Obtain the Sampling Stratum, SS, for each eligible cluster from the Sample Design File.

c. Obtain the PIL HU count, NHUILP for each eligible cluster from the Cluster Count File.

d. Calculate the following tallies for both medium (SS = 2) and large (SS = 3) clusters in each A.C.E. reduction stratum in the current state:

i. Total number of medium clusters in the listing sample, in the i^{th} A.C.E. reduction stratum, $NCLUSTM_i$
ii. Total number of large clusters in the listing sample, in the i^{th} A.C.E. reduction stratum, $NCLUSTL_i$
iii. Preliminary Listing HU count within medium clusters, in the i^{th} A.C.E. reduction stratum, $PILM_i$
iv. Preliminary Listing HU count within large clusters, in the i^{th} A.C.E. reduction stratum, $PILL_i$

2. Calculate the Initial Take-Everys

a. Calculate the target number of interview HUs for the i^{th} A.C.E. reduction stratum.

For reduction strata 1 - 15:

$$T_i = THU \times \frac{K_i \times \left(PILM_i + \frac{WTL}{WTM} PILL_i \right)}{\sum_{i \in (1 \text{to} 15)} K_i \times \left(PILM_i + \frac{WTL}{WTM} PILL_i \right)}$$

For reduction stratum 16:

$T_{16} = 1.0.$

b. Calculate the initial take-every for the medium clusters in the i^{th} A.C.E. reduction stratum.

For reduction strata 1 - 15:

$$ITEM_i = \frac{PILM_i + \frac{WTL}{WTM} \times PILL_i}{T_i}$$

If $ITEM_i < 1$, then output $ITEM_i$ to the updated Reduction Parameter File but continue the A.C.E. reduction process with $ITEM_i = 1.0$.

For reduction stratum 16:

$ITEM_{16}$ = TESJ as provided on the Reduction Parameter File.

c. Calculate the initial take-every for the large clusters in the i^{th} A.C.E. reduction stratum.

For reduction strata 1 - 15:

$$ITEL_i = ITEM_i \times \frac{LISTADJ}{TE2}$$

If $ITEL_i < 1.0$, then set $ITEL_i = 1.0$.

For convenience of future computations, set $ITEL_{16} = 1.0$ since stratum jumpers are either in the medium sampling stratum or the small sampling stratum and not in the large stratum.

3. Calculate the expected HU interview sample size and expected cluster interview sample size for both medium and large clusters in the i^{th} reduction stratum, including reduction stratum 16.

$$EHUM_i = \frac{PILM_i}{ITEM_i} \qquad ECLUSTM_i = \frac{NCLUSTM_i}{ITEM_i}$$

$$EHUL_i = \frac{PILL_i}{ITEL_i} \qquad ECLUSTL_i = \frac{NCLUSTL_i}{ITEL_i}$$

For stratum 16, let $EHUL_{16} = 0$ and $ECLUSTL_{16} = 0$.

Round $ECLUSTM_i$ and $ECLUSTL_i$ to integers using standard rounding. Denote these rounded numbers as follows:

$$RECLUSTM_i = int(ECLUSTM_i + 0.5)$$

$$RECLUSTL_i = int(ECLUSTL_i + 0.5)$$

If $NCLUSTM_i > 0$, then

If $RECLUSTM_i = 0$, then set $RECLUSTM_i = 1$ to ensure that one cluster is sampled in the stratum and continue to Step 4a;
If $RECLUSTM_i \neq 0$, then continue to Step 4a.

If $NCLUSTL_i > 0$, then

If $RECLUSTL_i = 0$, then set $RECLUSTL_i = 1$ to ensure that one cluster is sampled in the stratum and continue to Step 4c;
If $RECLUSTL_i \neq 0$, then continue to Step 4c.

4. Calculate the Final Take-Everys and Random Starts

a. Calculate the final take-every for the medium clusters in the i^{th} A.C.E. reduction stratum, including reduction stratum 16.

$$TEM_i = \frac{NCLUSTM_i}{RECLUSTM_i}$$

If $NCLUSTM_i = 0$, then set $TEM_i = 0$.

b. Generate a random number, RNM, between 0 and 1 ($0 < RNM \leq 1$), and calculate the random start for medium clusters in the i^{th} A.C.E. reduction stratum, including stratum 16 (medium stratum jumpers). Generate a new random number for each A.C.E. reduction stratum and each state.

$$RSM_i = TEM_i \times RNM$$

c. Calculate the final take-every for the large clusters in the i^{th} A.C.E. reduction stratum.

$$TEL_i = \frac{NCLUSTL_i}{RECLUSTL_i}$$

Set $TEL_{16} = 1.0$ since stratum jumpers are either in the medium sampling stratum or the small sampling stratum. TEL_{16} does not apply but is set for completeness.

If $NCLUSTL_i = 0$, then set $TEL_i = 0$.

d. Generate a random number, RNL, between 0 and 1 ($0 < RNL \leq 1$), and calculate the random start for large clusters in the i^{th} A.C.E. reduction stratum. Generate a new random number for each A.C.E. reduction stratum and each state.

$$RSL_i = TEL_i \times RNL$$

5. Calculate the cluster weight for medium and large clusters after the A.C.E. block cluster reduction for the i^{th} reduction stratum, including reduction stratum 16.

$$WTARM_i = WTM \times TEM_i$$
$$WTARL_i = WTL \times TEL_i$$

6. Update the Reduction Parameter File. The parameter file has one record for each reduction stratum per state. For each of the reduction strata, update the following variables:

Variable Description	Name	Location
Total medium clusters after the listing sample selection	NCLUSTM	86-90
Total large clusters after the listing sample selection	NCLUSTL	92-96
Preliminary Indep. Listing HU Count in medium clusters	PILM	98-103
Preliminary Indep. Listing HU Count in large clusters	PILL	105-110
Target interview sample size for the A.C.E. reduction stratum	T	111-120
Take-every for medium clusters	TEM	121-130
Take-every for large clusters	TEL	132-141
Random start for medium clusters	RSM	143-152
Random start for large clusters	RSL	154-163
Random number for medium clusters	RNM	165-172
Random number for large clusters	RNL	174-181
Expected number of housing units in medium clusters	EHUM	183-188
Expected number of housing units in large clusters	EHUL	190-195
Expected number of medium clusters	ECLUSTM	197-202
Expected number of large clusters	ECLUSTL	204-209
Medium cluster weight following A.C.E. reduction	WTARM	211-221
Large cluster weight following A.C.E. reduction	WTARL	223-233
Initial take-every for medium clusters	ITEM	234-242
Initial take-every for large clusters	ITEL	243-251

7. Provide the Reduction Parameter File to the Sample Design Team in the DSSD for review. If the calculation of the take-everys results in some values less than 1, then the differential sampling factors may need to be revised and the parameters recalculated. Wait for approval of the sampling parameters before proceeding to section F.

F. Select a Subsample of Block Clusters

For each of the A.C.E. reduction strata crossed with the original sampling strata, medium and large, select a separate systematic sample of block clusters as follows:

1. Sort the block clusters in the following order:

   - Sampling Stratum (SS).
   - A.C.E. Reduction Stratum (ARST).
   - Consistency Stratum (CONCODE).
   - List/Enumerate Indicator (LEIND).
   - Index Number (INDEX1) on the Sample Design File.

2. Assign an order number to each cluster in the sampling stratum and A.C.E. reduction stratum currently being subsampled. Give the first cluster in the sort an order number of "1", and increment by one for all remaining clusters. The assigned number is referred to as the A.C.E Reduction Index Number. Place the A.C.E. Reduction Index Number (INDEXR) on the Sample Design File.

3. Generate a sequence of numbers $L_1, \ldots, L_n$ as follows:

   - Obtain the Random Start (RSAR) and the Take-every (TEAR) for A.C.E. Reduction. If the current sampling stratum is medium (SS = 2), set RSAR = RSM and TEAR = TEM, where RSM and TEM are obtained from the Reduction Parameter File. If the current sampling stratum is large (SS = 3), set RSAR = RSL and TEAR = TEL, where RSL and TEL are obtained from the Reduction Parameter File.

   - Let L_1 = RSAR

   - Calculate $L_j = L_{j-1}$ + TEAR, for j = 2 to n where n is the largest integer such that [RSAR + (n - 1) × TEAR] ≤ N, where N is the largest order number in the sampling stratum and A.C.E. reduction stratum currently being subsampled.

   - Round each L_j up to the nearest integer (an integer rounds to itself).

- For each block cluster in the sampling stratum and the A.C.E. reduction stratum:

  If the order number is equal to the rounded values of L_j, j = 1, ..., n, then do the following:

  - Assign the A.C.E. Reduction Indicator (ACERED) on the Sample Design File equal to "1". The block cluster was selected in sample.

  - Calculate the block cluster weight (WEIGHTAR) following the A.C.E. Reduction. Obtain the Take-every for listing sample selection, TE1, and the second-step Take-every, TE2, from the Sample Design File, and compute the weight as follows;

    WEIGHTAR = TE1 × TE2 × TEAR

  If the order number does not equal any of the rounded values of L_j, j = 1, ..., n, then do the following:

  - Assign the A.C.E. Reduction Indicator (ACERED) on the Sample Design File equal to "0". The block cluster was not selected in sample.

  - Set the Current Sample Indicator (CSI) on the Sample Design File equal to "0". The block cluster was not selected so it is not currently in sample.

- For example: if N = 100, RSAR = 2.4 and TEAR = 7.2, then n = 14. Set L_1 = 2.4. The generated L_js would be the sequence: 2.4, 9.6, 16.8, 24.0, ..., 96.0. Therefore, the block clusters with ordered numbers 3, 10, 17, 24, 32, ..., and 96 would be selected for the sample.

4. Compute a Check

For each reduction stratum, check the number of sampled block clusters, given by n, by calculating c:

$$c = \left| \frac{N}{TEAR} - n \right|$$

If the sampling is implemented correctly, c will be less than 1. For values of c that are not less than one and have not been resolved, contact the DSSD for review of the sampling operations.

G. Update and Create Files

1. Update the Sample Design File. This file tracks the path that each sampled block cluster travels during the A.C.E. sampling procedures. It was created following the listing sample selection and contains one record per block cluster selected during the listing sample selection. Version 2 will be created by updating version 1 with the A.C.E. reduction information. The file layout is in Attachment D. Update the file with the following A.C.E. block cluster reduction information:

Variable Description	Name	Location
Preliminary Number of HUs on the Independent List	NHUILP	176-180
Number of HUs on the January 2000 DMAF	NHUDMAF	182-186
Demographic Code	DEMCODE	188
Consistency Code	CONCODE	189
A.C.E. Reduction Stratum (zero-filled)	ARST	190-191
A.C.E. Reduction Indicator	ACERED	193
Random Start for A.C.E. Reduction	RSAR	195-205
Take-every for A.C.E. Reduction	TEAR	207-217
Unbiased weight after A.C.E. reduction	WEIGHTAR	219-230
Collapsing Flag	COLFLAG	232
A.C.E. Reduction Index Number	INDEXR	234-241
Number of HUs on the December 1999 DMAF (Initial)	NHUDMAFI	243-247

2. Housing Unit Sample Size File. This file contains three variables which are the basis of the input to large block cluster subsampling. Create one record for each reduction stratum per state which includes the following variables:

Variable Description	Name	Location
State code	STATE	1-2
A.C.E. Reduction Stratum (zero-filled)	ARST	4-5
Target Number of Housing Units to Interview in A.C.E. Reduction Stratum	T	7-14

VIII REFERENCES

1 DSSD Census 2000 Procedures and Operations Memorandum Series R-8, "Census 2000 Specifications for Block Cluster Formation-Reissue," May 3, 1999.

2 DSSD Census 2000 Procedures and Operations Memorandum Series R-9, "Amendment to Census 2000 Specifications for Block Cluster Formation–Reissue," May 3, 1999.

3 DSSD Census 2000 Procedures and Operations Memorandum Series R-10, "Accuracy and Coverage Evaluation (ACE) Survey: Second Amendment to Census 2000 Specifications for Block Cluster Formation–Reissue," May 3, 1999.

4 DSSD Census 2000 Procedures and Operations Memorandum Series R-5, "Accuracy and Coverage Evaluation Survey: Universe File and Block Cluster Sampling Parameter File Specification," March 30, 1999.

5 DSSD Census 2000 Procedures and Operations Memorandum Series R-3, "Accuracy and Coverage Evaluation Survey: Block Cluster Sample Selection Specification," March 29, 1999.

6 DSSD Census 2000 Procedures and Operations Memorandum Series R-23, "Accuracy and Coverage Evaluation Survey: Approval and Summary of Results of the Reduction Sample," January 21, 2000.

7 DSSD Census 2000 Procedures and Operations Memorandum Series R-22, "Accuracy and Coverage Evaluation Survey: Cluster Reduction Contingency Plan," December 16, 1999.

8 DSSD Census 2000 Procedures and Operations Memorandum Series, Chapter S-FA-02 - Revision #1, TMO A.C.E. 2000 Planning Memorandum Series #2, "Revision #1 of A.C.E. 2000 Independent Listing File Transfers (Draft)," July 29, 1999–DRAFT

9 DSSD Census 2000 Procedures and Operations Memorandum Series D-1, "Specification of the Decennial Master Address File Deliverability Criteria for Census 2000," June 30, 1999.

cc: DSSD Census 2000 Procedures and Operations Memorandum Series Distribution List
A.C.E. Implementation Team
Statistical Design Team Leaders
DSSD Sample Design Team

A.C.E. Reduction Strata

Code	Stratum Name
01	Minority
02	Non-minority Low Inconsistent
03	Non-minority Consistent
04	Non-minority High Inconsistent
05	Non-minority Inconsistent
06	Non-minority
07	Low Inconsistent
08	Consistent
09	High Inconsistent
10	Inconsistent
11	Minority Inconsistent
12	Minority Consistent
13	Full Collapse
14	Minority Low Inconsistent
15	Minority High Inconsistent
16	Medium Stratum Jumpers
17	American Indian Reservations
18	Puerto Rico
19	Small Stratum Jumpers

A.C.E. Reduction Parameter File Layout and Example

Variable Description	Name	Location	Format
FIPS State Code	STATE	1-2	I2
A.C.E. Reduction Stratum	ARST	4-5	I2
State Target Number of HUs to Interview	THU	7-11	I5
Listing Adjustment for State	LISTADJ	13-20	F8.4
Second-step Take-every for state	TE2	22-29	F8.4
Medium Stratum Jumper Take-every	TESJ	31-38	F8.4
Low Inconsistency Cutoff	XL	40-43	F4.2
High Inconsistency Cutoff	XH	45-48	F4.2
Collapsing Flag	COLFLAG	50-51	I2
Differential Sampling Factor	K	53-61	F9.4
Medium cluster weight after the listing sample selection	WTM	63-72	F10.4
Large cluster weight after the listing sample selection	WTL	74-83	F10.4

Example of Reduction Parameter File for Alabama (State Code = 1):

```
         1         2         3         4         5         6         7         8
         0         0         0         0         0         0         0         0
01 01 4470  1.0000  1.0000  1.0000 0.75 1.25  1   2.0000  168.8000   34.4000
01 02 4470  1.0000  1.0000  1.0000 0.75 1.25  1   1.5000  168.8000   34.4000
01 03 4470  1.0000  1.0000  1.0000 0.75 1.25  1   1.0000  168.8000   34.4000
01 04 4470  1.0000  1.0000  1.0000 0.75 1.25  1   1.7500  168.8000   34.4000
```

In this example, the state total number of HUs to interview is 4,470, and both the listing adjustment and second-step take-every are 1. The Take-every for medium stratum jumpers is also 1. The cutoff for defining Low Inconsistency clusters is 0.75, meaning the PIL HU count must be at least 25 percent lower than the DMAF HU count for a cluster to be in the Low Inconsistent stratum. Likewise, the High Inconsistent Cutoff is 1.25, so the PIL must be at least 25 percent higher than the DMAF to be a High Inconsistent cluster. The collapsing flag is 1, meaning all minority clusters are in the same A.C.E. reduction stratum, while the non-minority clusters remain split into the three consistency strata, resulting in four A.C.E. reduction strata in Alabama and thus the four records in the example (See Table 5 on page 17). The A.C.E. reduction stratum codes are given in the second field. The differential sampling factor for each stratum is in the tenth field. The differential sampling factors give an indication of the differential sampling that occurs in Alabama in this example. Minority clusters will be sampled at twice the rate of non-minority consistent clusters, so the take-every for minority clusters is half the take-every of non-minority consistent clusters. Similarly, the take-every for non-minority low inconsistent clusters is 2/3 that of the non-minority consistent clusters. The medium cluster weight following the listing sample selection for Alabama is 168.8 and the large cluster weight is 34.4.

It is important to note that this is an example for illustrative purposes. It is very likely that the production parameters for Alabama will differ from those in this example.

Cluster Count File Layout

This file will be provided by the TMO.

Variable Description	Name	Location
LCO Number (digits 1-2 are RO code)	LCO	1-4
FIPS State Code	ST	5-6
FIPS County Code	CC	7-9
Cluster Number (with check digit)	CLUSTER	10-15
Preliminary number of Independent Listing HUs	NHUILP	16-20

Sample Design File Layout

Variable Description	Name	Places	Source
Census Region	REGION	1	UN
Census Division	DIV	2	UN
State code	STATE	3-4	UN
County code	COUNTY	5-7	UN
Local census office	LCO	8-11	CS
Interim Tract (Pseudo Tract)	ITRACT	12-17	BC
Current Sample Indicator	CSI	19	UO
A.C.E. block cluster number	CLUST	21-25	CS
Check Digit	DIGIT	26	CS
Geography block cluster number	GCLUST	28-32	BC
List/Enumerate Indicator	LEIND	33	BC
Type of Enumeration Area Recode	TEACR	34	CS
Type of Enumeration Area group	TEAG	36	BC
Number of HUs used for sample design	NHU	37-41	BC
Number of MAF HUs	NHUM	43-47	BC
Number of 1990 HUs	NHU90	49-53	BC
Sampling Stratum 1 = Small 2 = Medium 3 = Large 4 = American Indian Reservation	SS	55	UN
American Indian Country Indicator 0 = No American Indian Country 1 = American Indian Reservation/trust land 2 = Tribal Jurisdiction Area/ Alaska Native Village Statistical Area/ Tribal Designated Statistical Area	AICIND	56	BC
Demographic/Tenure Group code	DTCODE	57-58	UN
Demographic/Tenure Group label	DTLABEL	59-60	UN
Estimated Urbanicity of block cluster 1 = Urban Area with population ≥250,000 2 = Other Urban Area 3 = Non-Urban Area	ECLUSURB	62	UN
Size Category 1 = Small (0-2 hus) 2 = Medium (3-79 hus) 3 = Large (80+ hus)	SIZCAT	63	UN
Additional space		64-91	
First step index number	INDEX1	92-99	CS
Listing sample selection Indicator 1 = Selected	BC1	101	CS
Random Start for listing sample selection	RS1	103-113	UN
Take-every for listing sample selection	TE1	115-125	UN
Second block cluster sampling Indicator 0 = Not Selected 1 = Selected	BC2	127	CS

Random Start for second block cluster sampling	RS2	129-139	CS
Take-every for second block cluster sampling	TE2	141-151	CS
Unbiased weight after block cluster sampling	WEIGHTBC	153-164	CS
Additional space		165-175	
Preliminary Number of HUs on the Independent List	NHUILP	176-180	AR
Number of Housing Units on the January 2000 DMAF	NHUDMAF	182-186	AR
Demographic Code 1 = Minority 2 = Non-minority 3 = Puerto Rico	DEMCODE	188	AR
Consistency Code 1 = Low Inconsistent (PIL significantly smaller than DMAF) 2 = Consistent 3 = High Inconsistent (PIL significantly larger than DMAF)	CONCODE	189	AR
A.C.E. Reduction Stratum (zero-filled)	ARST	190-191	AR
A.C.E. Reduction Indicator 0 = Not Selected 1 = Selected	ACERED	193	AR
Random Start for A.C.E. Reduction	RSAR	195-205	AR
Take-every for A.C.E. Reduction	TEAR	207-217	AR
Unbiased weight after A.C.E. reduction	WEIGHTAR	219-230	AR
Collapsing Flag	COLFLAG	232	AR
A.C.E. Reduction Index Number	INDEXR	234-241	AR
Number of Housing Units on the December 1999 DMAF (Initial)	NHUDMAFI	243-247	AR
Additional space		248-300	

Source Codes

AR: A.C.E. Reduction
BC: Block Clustering
CS: Block Cluster Sampling
UN: Universe File Creation
UO: Updated for each operation

UNITED STATES DEPARTMENT OF COMMERCE
Bureau of the Census
Washington, DC 20233-0001

May 16, 2000

DSSD CENSUS 2000 PROCEDURES AND OPERATIONS MEMORANDUM SERIES #R-31

MEMORANDUM FOR Maureen Lynch
Assistant Division Chief, Coverage Measurement Processing
Decennial Statistical Studies Division

From: Donna Kostanich DK
Assistant Division Chief, Sampling and Estimation
Decennial Statistical Studies Division

Prepared by: Ryan Cromar RC, James Farber JF, and Roxanne Feldpausch RF
Decennial Statistical Studies Division

Subject: Accuracy and Coverage Evaluation Survey: Specification for E-Sample Identification

I. INTRODUCTION

This specification describes the identification of E-sample housing units (HUs) for the Accuracy and Coverage Evaluation (A.C.E.) survey. The identification of the E sample is required for person matching and final HU matching. In general, E-sample HUs are HUs enumerated in the census in block clusters or segments of clusters that were in the P sample. This specification covers only the identification of E-sample HUs. The E-sample persons are all data-defined persons enumerated in the census in E-sample HUs, and are identified when the person matching files are created.

The E sample is used primarily for A.C.E. estimation. The E sample and P sample will be matched to estimate how many persons and HUs were omitted from or erroneously included in the census. The P sample was created through the A.C.E. sampling process, including listing sample selection, the A.C.E. reduction, small block cluster subsampling, and within-large block cluster subsampling, and consists of all HUs and persons interviewed in the independent A.C.E. sample. P-sample HUs are contained on the subsampled preliminary enhanced list (SPEL), an output of large block cluster subsampling.

The A.C.E. HUs were originally identified through the independent listing operation and placed on the independent list. Census HUs were identified through a number of census operations and were placed on the Decennial Master Address File (DMAF). During initial HU matching and follow-up, the two address lists were compared and the results recorded to the preliminary enhanced list. The preliminary enhanced list contains the A.C.E. and census HUs that match to each other along with HUs from each list that do not match but were found to be valid HUs during field follow-up. The SPEL is the preliminary enhanced list updated with the results of large block cluster subsampling. The same HUs are on both the preliminary enhanced list and the SPEL. The P sample consists of A.C.E. HUs on the SPEL in block clusters or segments of clusters selected for interview. Census HUs may also be on the SPEL, but they are ineligible for interview.

The E sample is identified to maximize the overlap between the E sample and P sample. The overlapping is not required for A.C.E. estimation, but it increases efficiency by reducing field follow-up workloads. Field follow-up is also reduced by the E-sample subsampling described in this specification. When the number of E-sample eligible HUs in a block cluster is too large, a subsample of these HUs will be selected for inclusion in the E sample.

This specification is ordered into the following sections:

- Definitions
- Assumptions
- Overview
- Input Files
- Output Files
- Process
- Verification
- References

This specification should be used to flowchart the process, to generate further discussion on requirements, to identify and finalize the record layouts of input and output files, and to write computer software to implement the methodology. During and after a testing phase, it is possible that changes to the specification will be necessary.

Any questions or comments regarding this specification should be directed to Ryan Cromar (301-457-1636), James Farber (301-457-4282), or Deborah Fenstermaker (301-457-4195) of the Decennial Statistical Studies Division (DSSD).

II. DEFINITIONS

A. A.C.E. Housing Unit

A housing unit on the SPEL with an after follow-up match code (see below) of M, MU, CI, or UI. In general, these are HUs found during A.C.E. independent listing.

B. After Follow-up Match Codes

Codes assigned to HUs during initial A.C.E. HU matching and follow-up. For the purposes of this specification, the only match codes that need to be defined are those that occur on the SPEL. As documented in reference 1, these match codes are:

M	=	The A.C.E. and census addresses match.
MU	=	The A.C.E. and census addresses match and there is not enough information on the follow-up form to confirm this match as an HU with certainty. The follow-up interview was not done, was incomplete, was never sent, had contradictory information, or was a noninterview.
UI	=	Not enough information on the follow-up form to assign a code to the nonmatched A.C.E. HU with certainty. The follow-up interview was not done, was incomplete, was never sent, had contradictory information, or was a noninterview.
UE	=	Not enough information on the follow-up form to assign a code to the census nonmatched HU with certainty. The follow-up interview was not done, was incomplete, was never sent, had contradictory information, or was a noninterview.
CI	=	The A.C.E. HU existed as an HU at the time of the follow-up interview and is correctly geocoded in the block cluster. The HU is not found in the census.
CE	=	The census HU existed as an HU at the time of the follow-up interview and is correctly geocoded in the block cluster. The HU is not found in the A.C.E.

C. American Indian Reservation Block Cluster

A block cluster at least partially on an American Indian Reservation (AIR), according to the AIR definitions used at the time of A.C.E. block clustering. The AIR block clusters have an American Indian Country Indicator (AICIND) = 1 on the Sample Design File.

D. Corresponding HU

A census HU on the E-Sample Identification Input File that corresponds to an SPEL HU. Census HUs correspond to SPEL HUs through the Census Identification Number (CID), a code assigned to all census HUs. During initial HU matching and follow-up, SPEL HUs that were matched to HUs on the DMAF were assigned the CID of the matching DMAF HU. If the DMAF HU was not deleted in later operations, such as nonresponse follow-up, and made it onto the Hundred Percent Census Unedited File (HCUF), the source of the E sample, then the CID link still exists between the SPEL HU and the HCUF HU. The SPEL HU has the same CID as the corresponding HCUF HU. Since both A.C.E. and supplemental HUs are on the SPEL, an HCUF HU may correspond to either type of SPEL HU.

The correspondence status of an HCUF HU is denoted by the variable Match Status (MSTATUS), where

0	=	The HCUF HU is in a cluster with fewer than 80 HCUF HUs.
1	=	The HCUF HU is corresponding in a cluster with 80 or more HCUF HUs.
2	=	The HCUF HU is non-corresponding in a non-special case cluster with 80 or more HCUF HUs.
3	=	The HCUF HU is non-corresponding and in a special case cluster with 80 or more HCUF HUs.

E. E Sample

Census HUs in block clusters or block cluster segments that were selected for the A.C.E. interview sample. The E sample can also refer to the persons enumerated in the census in E-sample HUs. The HCUF is the source of the E sample.

F. E-Sample Indicator

Indicates whether or not an HCUF HU is in the E sample.

1	=	The HCUF HU is in the E sample.
2	=	The HCUF HU is not in the E sample.

G. E-Sample Probability Code

Indicates which E-sample weight applies to an E-sample HU.

1	=	WEIGHTE1 applies to the E-sample HU.
2	=	WEIGHTE2 applies to the E-sample HU.

It is possible for two weights to apply to the E sample in a single block cluster if E-sample subsampling is required in the cluster.

H. HU Group

Indicates how to calculate the weight for each E-sample HU and in which set of summary HU counts each E-sample HU should be tallied. Attachment A contains a description of the HU Groups.

I. Hundred Percent Census Unedited File Housing Unit

An HU found to be a valid HU during the census and thus retained from the DMAF onto the HCUF. The HUs on the E-Sample Identification Input File are also called HCUF HUs since the E-sample input file is simply an extract of the HCUF.

J. Non-Corresponding Housing Unit

A census HU on the E-Sample Identification Input File whose CID was not assigned to any SPEL HUs during initial HU matching. Examples of non-corresponding HCUF HUs include census adds and formerly mis-geocoded HUs that are moved into the cluster.

K. P sample

HUs in the A.C.E. interview sample. The P sample can also refer to the persons interviewed in the A.C.E. in P-sample HUs. The P-sample HUs were initially identified during A.C.E. independent listing.

L. Preliminary E-Sample Indicator

Indicates whether or not an HCUF HU is initially eligible for the E sample. Some HCUF HUs may initially be eligible for the E sample but will not end up in the final E sample due to E-sample subsampling.

0	=	The HCUF HU is not initially eligible for the E sample and is not subject to E-sample subsampling.
1	=	The HCUF HU is initially eligible for the E sample and may be subject to E-sample subsampling.

M. Segment Identifier

A variable created during large block cluster subsampling and mapped onto HCUF HUs. The field assignment segments in the control number are not the segment identifiers referred to in this specification. See reference 2 for information on how segment identifiers were created in large block cluster subsampling.

N. Supplemental HU

A housing unit on the SPEL with an after follow-up match code of UE or CE. Supplemental HUs are not eligible for the A.C.E. interview sample, but were assigned segment identifiers and special interview codes during large block cluster subsampling to facilitate E-sample identification. The source of supplemental HUs was the version of the DMAF used for initial HU matching. Supplemental HUs were not matched to any A.C.E. HUs during initial HU matching but were found to exist during HU follow-up.

III. ASSUMPTIONS

A. Supplemental HUs were map spotted during initial HU follow-up and thus can be correctly located near their geographic neighbors on the SPEL.

B. HCUF HUs correspond to SPEL HUs through the CID. The CIDs of corresponding HCUF HUs were assigned to their matching SPEL HUs during initial HU matching.

C. There is no E-sample subsampling in AIR clusters, using the AIR definitions known at the time of A.C.E. block clustering.

D. The E-sample persons will be identified during the creation of files for A.C.E. person matching and are not within the scope of this specification. The E-sample persons are all data-defined persons on the HCUF in E-sample HUs (see reference 3).

E. All decimal numbers are rounded to six decimal places at the time of creation using standard rounding procedures, unless otherwise noted. In this specification, standard rounding means that a number with a seventh-decimal value of five or higher is rounded up in the sixth decimal. Otherwise, the sixth decimal value is unchanged.

IV. PROCESS OVERVIEW

The E-sample identification process described in this specification is used to identify which census HUs are in the E sample. The E-sample HUs and the persons in those HUs will be matched to the HUs and persons in the P sample, the A.C.E. interview sample. Those matching results form the basis of A.C.E. person and final HU estimation. This overview details the E-sample identification process. The steps in this section correspond to the steps in section VII, which contains the programming instructions and is significantly less detailed than this overview. Attachment B contains a flowchart of the E-sample identification process.

A. The first step is to screen out clusters where the identification of the E sample is simple. The source of E-sample HUs is an extract of the HCUF, the E-Sample Identification Input File, which contains all census HUs in A.C.E. sample block clusters. The HCUF HUs not in A.C.E. clusters are excluded from the E-Sample Identification Input File, as are all group quarters records and all person records.[1]

   If the number of HCUF HUs in a cluster is less than 80, then all of those HUs are in the E sample. Likewise, if a cluster is an AIR cluster, using the AIR definitions known at the time of A.C.E. block clustering, then all HCUF HUs are in the E sample regardless of their number. For these clusters, E-sample identification is simple. Information for these clusters is saved to the E-Sample Identification Input File and the Sample Design File, and they are finished with the E-sample identification process. Clusters that have at least 80 HCUF HUs and are not on an AIR continue to the next step.

B. There are four types of clusters that are considered special cases in the E-sample identification process because all of their HCUF HUs are non-corresponding. A

---

[1]In this specification, HUs on the E-Sample Identification Input File are also called HCUF HUs.

non-corresponding HU is an HCUF HU whose CID was not assigned to an SPEL HU during initial HU matching. For example, all HCUF adds are non-corresponding. The four types of special case clusters are:

- List/Enumerate clusters, which did not go through initial HU matching since they did not have any DMAF HUs at that point.
- Relisted clusters, which also did not go through initial HU matching due to timing constraints.
- Clusters that went through initial HU matching and had no SPEL HUs that were assigned a CID. These are clusters where all of the SPEL HUs have match codes of UI or CI.
- Clusters that had zero A.C.E. HUs and zero supplemental HUs in initial HU matching but have 80 or more HCUF HUs.

All HCUF HUs in special case clusters are initially eligible for the E sample. In special case clusters with at least 80 HCUF HUs, E-sample subsampling is required, and these clusters proceed immediately to step E. (Special case clusters with fewer than 80 HUs were screened out in step A.)

C. In non-special case clusters with at least 80 HCUF HUs, the next step is to map the results of large block cluster subsampling onto the HCUF and identify those HUs that are in the E sample, out of the E sample, or have an unknown E-sample status. Corresponding HUs are HCUF HUs whose CIDs were assigned to SPEL HUs during initial HU matching. The E-sample status of corresponding HUs can be fully determined in this step using the results of large block cluster subsampling. Non-corresponding HUs have an unknown E-sample status at this point.

- If large block cluster subsampling did not occur in the cluster, then all corresponding HCUF HUs are in the E sample.
- If large block cluster subsampling did occur in the cluster, then corresponding HCUF HUs in segments selected for the P sample are in the E sample. Corresponding HCUF HUs in segments not selected for the P sample are not in the E sample.
- For non-corresponding HCUF HUs, step D and possibly step E will determine which are in the E sample.

D. This step determines whether E-sample subsampling is required in a non-special case block cluster. Non-corresponding HUs were not available for large block cluster subsampling and thus received no segment identifiers. To determine whether E-sample subsampling is necessary, the first step is to assign segment identifiers to non-corresponding HUs. All HCUF HUs are sorted by geography, and non-corresponding HUs are assigned the segment identifier of the nearest previous corresponding HU.

Then, non-corresponding HUs in segments not selected for the P sample are out of the E sample. Non-corresponding HUs in P-sample segments are initially eligible for the E sample, but may be subsampled if there are 80 or more of such HUs. If there are fewer than 80 E-sample eligible non-corresponding HUs, then all of them are in the E sample. If there are 80 or more, then a subsample is drawn in step E.

E. Only clusters with 80 or more E-sample eligible HUs, including special case clusters, go through E-sample subsampling. The E-sample subsampling process is a standard systematic sample with a random start and a geographic sort of HCUF HUs. The take-every is the number of E-sample eligible HUs divided by 40, with a maximum take-every of 4. The E-sample eligible HUs that are selected in the subsample are in the E sample, while non-selected HUs are out of the E sample.

F. The last step of E-sample identification is to update files and compute summary counts. Note that the E-sample HUs within a single cluster may have different E-sample weights. If E-sample subsampling did not occur in a cluster, then the E-sample HUs in that cluster have only one weight. However, two E-sample weights may be required in clusters that were subsampled. In non-special case clusters where E-sample subsampling occurred, corresponding HUs have a different weight than non-corresponding HUs since only the latter were subject to E-sample subsampling and thus received additional weight. In special case clusters, only one weight applies to all E-sample HUs even if E-sample subsampling occurred since all HUs in special case clusters with 80 or more HUs are subject to E-sample subsampling.

V. INPUT FILES

A. Subsampled Preliminary Enhanced List

Description:	The SPEL is the preliminary enhanced list updated with the results of large block cluster subsampling. All HUs with the after follow-up match codes given in Section III.A are on the SPEL whether or not they were in segments selected for the P sample. There is one SPEL for each A.C.E. regional office.
Level:	Housing Unit
Scope:	One record for each HU in A.C.E. sample block clusters following small block cluster subsampling.
Layout:	See Attachment C.

B. E-Sample Identification Input File

Description:	The E-Sample Identification Input File is an extract of the HCUF, the file that contains the data for all HUs, group quarters, and data-defined persons enumerated in the census. The E-Sample Identification Input File includes certain variables from HCUF HU records but excludes person and group quarters records, and limits the geographic scope only to A.C.E. sample clusters. There is one E-Sample Identification Input File for each Local Census Office.
Level:	Housing Unit
Scope:	One record for each census HU in A.C.E. sample clusters.
Layout:	Not yet available.

C. Sample Design File Version 5

Description:	Version 5 of the A.C.E. Sample Design File, which reflects the previous A.C.E. sampling operations: listing sample selection, A.C.E. reduction, small block cluster subsampling, large block cluster subsampling, and targeted extended search sampling. There are 29,717 records on the Sample Design File. The name of version 5 is SDF.US5.
Level:	Block Cluster
Scope:	One record for each block cluster selected in the first step of the A.C.E. listing sample.
File Layout:	See Attachment D.

VI. OUTPUT FILES

A. Sample Design File Version 6

Description:	Version 6 of the A.C.E. Sample Design File, which includes the results of E-sample identification. Version 6 will be named SDF.US6.
Level:	Block Cluster
Scope:	One record for each block cluster selected in the first step of the A.C.E. listing sample.
File Layout:	See Attachment D.

B. Updated E-Sample Identification Input File

Description: The input E-Sample Identification Input File is updated with the results of E-sample identification. HCUF HUs are assigned

	indicators denoting whether they are in or out of the E sample, whether or not they were initially eligible for the E sample, which weight should apply to each HU, to which segment and HU group each HU was assigned, and whether they are corresponding or not.
Level:	Housing Unit
Scope:	One record for each census HU in A.C.E. sample clusters.
Layout:	Not yet available.

VII. PROCESS

Apply the following steps to each block cluster on the E-Sample Identification Input File. Attachment B contains a flowchart of the E-sample identification process, and Attachment E gives a summary table of the process.

A. Determine the Number of HCUF HUs in the Block Cluster

1. Tally the number of HCUF HUs in the block cluster, and denote this tally NHUCUF.

2. Obtain the American Indian Country Indicator (AICIND) for the cluster from the Sample Design File.

    - If NHUCUF $\geq$ 80 and AICIND $\neq$ 1, then proceed to step B below.

    - If NHUCUF < 80 or AICIND = 1, then assign the following variables to each HCUF HU in the cluster and proceed to step F where weights, summary counts, and other information will be determined.

        a. E-Sample Indicator = 1
        b. Preliminary E-sample Indicator = 0
        c. E-Sample Probability Code = 1
        d. Segment Identifier = AA
        e. HU Group = 1
        f. MSTATUS = 0

B. Determine if the Block Cluster is a Special Case Cluster

1. Obtain LEIND and RELIST for the block cluster from the Sample Design File.

2. Tally the following counts for the cluster and denote as indicated:

   a. Number of SPEL HUs in the cluster, NHUEL
   b. Number of SPEL HUs with after follow-up match codes of UI or CI, NUICI

3. If LEIND = 1, RELIST = 1, NHUEL = 0, or NHUEL = NUICI, then the block cluster is a special case cluster and continues to step 4.

   Otherwise, the cluster is not a special case cluster. Proceed to step C below.

4. Tally the number of HCUF HUs in the cluster, and denote this tally NHUCUFS2.

5. Assign the following variables to each HCUF HU in the cluster:

   a. Preliminary E-Sample Indicator = 1
   b. E-Sample Probability Code = 2
   c. Segment Identifier = AA
   d. MSTATUS = 3

6. Proceed to step E below since these clusters require E-sample subsampling. Special case clusters with fewer than 80 HCUF HUs were screened out in step A.

C. Determine the E-Sample Status for HCUF HUs

1. Compare the CIDs of HCUF HUs to the CIDs assigned to SPEL HUs in the cluster. Corresponding HCUF HUs have their CIDs represented on the SPEL, while non-corresponding HCUF HUs do not.

   - For corresponding HCUF HUs, set MSTATUS = 1

   - For non-corresponding HCUF HUs, set MSTATUS = 2

2. Obtain INTERVW and DSSDSEG from the SPEL for each corresponding HU.

3. Do the following for each corresponding HCUF HU in the cluster:

   a. Assign E-Sample Probability Code = 1

b. If INTERVW = 1 or 9, then assign the following variables:

   i. E-Sample Indicator = 1
   ii. Preliminary E-Sample Indicator = 0
   iii. Segment Identifier = DSSDSEG
   iv. HU Group = 2

c. If INTERVW = 0 or 8, then assign the following variables:

   i. E-Sample Indicator = 2
   ii. Preliminary E-Sample Indicator = 0
   iii. Segment Identifier = DSSDSEG
   iv. HU Group = 3

4. Do the following for each non-corresponding HCUF HU in the cluster:

   a. Assign E-Sample Probability Code = 2

D. Determine if E-Sample Subsampling is Required in Non-Special Case Clusters

Tally the number of non-corresponding HCUF HUs in the block cluster. If this tally is zero, then E-sample subsampling is not required; proceed to step F below. Otherwise, do the following to determine if subsampling is required:

1. Obtain TEACR from the Sample Design File. Sort the non-corresponding HCUF HUs with the corresponding HCUF HUs in the cluster as follows:

   - For clusters with city-style addresses (TEACR = 1), sort by:
     - Block and block suffix
     - Street name
     - House number
     - Unit designation
     - CID
   - For clusters with non-city-style addresses (TEACR = 2), sort by:
     - Block and block suffix
     - Census map spot number
     - Within map spot number
     - CID

2. Assign to each non-corresponding HCUF HU in the cluster the segment identifier of the corresponding HCUF HU prior to it. If the first HCUF HU in the sort is non-corresponding, assign it the segment identifier of the last segment assigned to that block cluster during large block cluster subsampling. See Attachment F for an example of this assignment.

3. Set the E-sample indicators for non-corresponding HCUF HUs as follows:

   - If INTERVW = 1 or 9 for HUs in the same segment on the SPEL, then assign the following variable to each non-corresponding HCUF HU in the cluster:

     a. Preliminary E-Sample Indicator = 1

   - If INTERVW = 0 or 8 for HUs in the same segment on the SPEL, then assign the following variables for each non-corresponding HCUF HU in the cluster:

     a. E-Sample Indicator = 2
     b. Preliminary E-Sample Indicator = 0
     c. HU Group = 7

4. Determine if E-sample subsampling is required in the cluster as follows:

   - Tally the number of HCUF HUs with a Preliminary E-Sample Indicator = 1 and denote this tally NHUCUFS2.

   - If NHUCUFS2 < 80 in the cluster then:

     a. For HCUF HUs with Preliminary E-Sample Indicator = 1, assign E-Sample Indicator = 1
     b. Assign the same HCUF HUs HU Group = 4
     c. Proceed to step F below.

   - If NHUCUFS2 ≥ 80 proceed to step E below.

E. E-sample Subsampling

Select a subsample of HCUF HUs with Preliminary E-Sample Indicator = 1 in a cluster as follows:

1. HUs in special case clusters have not yet been sorted. Obtain TEACR for the cluster from the Sample Design File. Sort the HCUF HUs with a Preliminary E-Sample Indicator = 1 by:

   - For clusters with city-style addresses (TEACR = 1), sort by:

     - Block and block suffix
     - Street name
     - House number
     - Unit designation
     - CID

   - For clusters with non-city-style addresses (TEACR = 2), sort by:

     - Block and block suffix
     - Census map spot number
     - Within map spot number
     - CID

2. Calculate the take-every, TEES:

   - $$TEES = \frac{NHUCUFS2}{40}$$

   - If TEES > 4.000000, set TEES = 4.000000.

3. Assign each HCUF HU with Preliminary E-Sample Indicator = 1 an order number, ON, starting at one and incrementing by one until all such HUs in the cluster have an ON. The largest order number will equal the value of NHUCUFS2.

4. Generate a sequence of numbers $L_1, ..., L_n$ as follows:

   a. Generate a random number between 0 and 1 ($0 < RN \leq 1$).

   b. Calculate a random start, $RSES = RN \times TEES$.

   c. Let $L_1 = RSES$.

   d. Calculate $L_j = L_{j-1} + TEES$, for $j = 2, ..., n$, where n is the largest integer such that $[RSES + (n - 1) \times TEES] \leq NHUCUFS2$.

   e. Round each L_j up to the nearest integer (an integer rounds to itself).

f. Each HCUF HU with Preliminary E-Sample Indicator = 1 and with ON equal to the rounded values of L_j, $j = 1, ..., n$, is in the E sample. Assign the following for each of these HUs:

   i. E-Sample Indicator = 1
   ii. For HUs in non-special case clusters, set HU Group = 5
   iii. For HUs in special case clusters, set HU Group = 8

g. Each HU with Preliminary E-Sample Indicator = 1 and with ON not equal to the rounded values of L_j, $j = 1, ..., n$, is not in the E sample. Assign the following for each of these HUs:

   i. E-Sample Indicator = 2
   ii. For HUs in non-special case clusters, set HU Group = 6
   iii. For HUs in special case clusters, set HU Group = 9

For example, let NHUCUFS2 = 122 and RN = 0.345167. Then TEES = 3.050000, RSES = 1.052759 and n = 40. Set L_1 = 1.052759. The generated L_js would be the sequence: 1.052759, 4.102759, 7.152759, 10.202759, 13.252759, 16.302759, 19.352759, 22.402759, 25.452759, 28.502759, 31.552759, 34.602759, 37.652759, 40.702759, 43.752759, 46.802759, 49.852759, 52.902759, 55.952759, 59.002759, 62.052759, 65.102759, 68.152759, 71.202759, 74.252759, 77.302759, 80.352759, 83.402759, 86.452759, 89.502759, 92.552759, 95.602759, 98.652759, 101.702759, 104.752759, 107.802759, 110.852759, 113.902759, 116.952759 and 120.002759. Therefore, the HUs with ON values of 2, 5, 8, 11, 14, 17, 20, 23, 26, 29, 32, 35, 38, 41, 44, 47, 50, 53, 56, 60, 63, 66, 69, 72, 75, 78, 81, 84, 87, 90, 93, 96, 99, 102, 105, 108, 111, 114, 117, and 121 would be selected for the sample.

5. Check the number of sampled HUs by calculating c:

$$c = \left| \frac{\text{NHUCUFS2}}{\text{TEES}} - n \right|$$

If the sampling is implemented correctly, c will be less than 1. For values of c that are greater than or equal to one, contact the Sample Design Team in the DSSD for review of the sampling operations.

F. Update Files

1. Sample Design File Version 6

a. E-sample Weights

On the Sample Design File, there will be two cluster-level E-sample weights, WEIGHTE1 and WEIGHTE2. HCUF HUs with an E-Sample Probability Code = 1 have a weight of WEIGHTE1. HCUF HUs with an E-Sample Probability Code = 2 have a weight of WEIGHTE2. If a cluster has only one type of HU Group, then only one weight variable will apply and the other weight will remain blank.

Record weights using the following criteria:

i. If a cluster has HCUF HUs with HU Group = 1, set WEIGHTE1 = WEIGHTC, which is obtained from the Sample Design File Version 5.

ii. If a cluster has HCUF HUs with HU Group = 2, set WEIGHTE1 = WEIGHTP from the Sample Design File Version 5.

iii. If a cluster has HCUF HUs with HU Group = 4, set WEIGHTE2 = WEIGHTP.

iv. If a cluster has HCUF HUs with HU Group = 5, set WEIGHTE2 = TE1×TE2×TEAR×FTESB×TELB×TEES, where TE1, TE2, TEAR, FTESB, and TELB are obtained from the Sample Design File.

v. If a cluster has HCUF HUs with HU Group = 8, set WEIGHTE2 = TE1×TE2×TEAR×FTESB×TEES.

b. HCUF and E-sample HU Counts

Compute nine HU counts for each cluster according to Table 1 below and record these totals to the Sample Design File. A value of 1 in the HU Group column in Table 1 indicates the counts in which an HCUF HU in that HU Group should be included. For example, an HCUF HU falling into HU group 1 is included in NHUES1, NHUES, NHUCUFS1, NHUCUFS, NHUCUF1, and

NHUCUF. However, an HCUF HU falling into HU group 3 is only included in NHUCUF1 and NHUCUF.

Table 1. HCUF and E-Sample Housing Unit Counts

Count	HU Group								
	1	2	3	4	5	6	7	8	9
NHUES1	1	1	0	0	0	0	0	0	0
NHUES2	0	0	0	1	1	0	0	1	0
NHUES	1	1	0	1	1	0	0	1	0
NHUCUFS1	1	1	0	0	0	0	0	0	0
NHUCUFS2	0	0	0	1	1	1	0	1	1
NHUCUFS	1	1	0	1	1	1	0	1	1
NHUCUF1	1	1	1	0	0	0	0	0	0
NHUCUF2	0	0	0	1	1	1	1	1	1
NHUCUF	1	1	1	1	1	1	1	1	1

Notes:

1. In clusters with fewer than 80 HCUF HUs, NHUCUF1, NHUCUF, NHUCUFS1, NHUCUFS, NHUES1, and NHUES are all equal since all HCUF HUs are in the E sample. In addition, NHUCUF2, NHUCUFS2, and NHUES2 are always equal to zero in these clusters since there are no HCUF HUs with an E-Sample Probability Code of 2.
2. In clusters with 80 or more HCUF HUs, NHUCUFS1 and NHUES1 are equal since all corresponding HCUF HUs in selected segments are in the E sample.
3. For special case clusters with 80 or more HCUF HUs, NHUCUF2 and NHUCUFS2 are equal since no segment boundaries can be defined. In addition, NHUCUF1 and NHUCUFS1 are always equal to zero in these clusters since there are no HCUF HUs with an E-Sample Probability Code of 1.

NHUES1, NHUES2, and NHUES are the number of HCUF HUs in the E sample with each E-Sample Probability Code.

- NHUES1: the number of E-sample HUs with E-Sample Probability Code = 1
- NHUES2: the number of E-sample HUs with E-Sample Probability Code = 2
- NHUES: the total number of E-sample HUs

NHUCUFS1, NHUCUFS2, and NHUCUFS are the number of HCUF HUs in segments selected for the P sample with each

E-Sample Probability Code.

- NHUCUFS1: the number of HCUF HUs in selected segments with E-Sample Probability Code = 1
- NHUCUFS2: the number of HCUF HUs in selected segments with E-Sample Probability Code = 2
- NHUCUFS: the total number of HCUF HUs in selected segments

NHUCUF1, NHUCUF2, and NHUCUF are the number of HCUF HUs in the cluster with each E-Sample Probability Code.

- NHUCUF1: the number of HCUF HUs with E-Sample Probability Code = 1.
- NHUCUF2: the number of HCUF HUs with E-Sample Probability Code = 2.
- NHUCUF: the total number of HCUF HUs in the block cluster.

c. E-Sample Identification Cluster Category

Set the E-Sample Identification Cluster Category (EICC) on the Sample Design File for each cluster using the following rules:

- If NHUCUF < 80 then EICC = 1
- If NHUCUF ≥ 80 and NHUCUFS < 80 then EICC = 2
- If NHUCUF ≥ 80 and NHUCUFS ≥ 80 then EICC = 3
- If NHUCUF ≥ 80 and RELIST = 1 then EICC = 4
- If NHUCUF ≥ 80 and LEIND = 1 then EICC = 5
- If NHUCUF ≥ 80 and NHUEL > 0 and NHUEL = NUICI then EICC = 6
- If NHUCUF ≥ 80 and NHUEL = 0 then EICC = 7

d. E-Sample Subsampling Information

Record the following variables for each cluster. If E-sample subsampling was not required in a cluster, set both variables equal to one.

- E-sample subsampling random start, RSES
- E-sample subsampling take-every, TEES

2. E-Sample Identification Input File

Update the E-Sample Identification Input File with the results of E-sample identification. Specifically, record the following variables for each HCUF HU:

- E-Sample Indicator
- Preliminary E-Sample Indicator
- E-Sample Probability Code
- Segment Identifier
- HU Group
- MSTATUS, the HCUF HU correspondence indicator

The layout of this file and thus the locations of these variables is not yet known.

VIII. VERIFICATION

Verification of the E-sample identification will include both micro-level independent replication of the process and macro-level analysis of to-be-determined summary statistics. The Sample Design Team in the DSSD will perform all activities involved in both types of verification. Access to the following files is required:

- Sample Design File Version 5
- Sample Design File Version 6
- Subsampled Preliminary Enhanced List
- Updated E-Sample Identification Input File

IX. REFERENCES

1 DSSD Census 2000 Procedures and Operations Memorandum Series, Chapter S-HU-08, "Creation of the Census 2000 Accuracy and Coverage Evaluation (A.C.E.) Enhanced List for Person Phase Interviewing," June 21, 1999, DRAFT.

2 DSSD Census 2000 Procedures and Operations Memorandum Series R-27, "Accuracy and Coverage Evaluation: Large Block Cluster Subsampling Specifications," March 8, 2000.

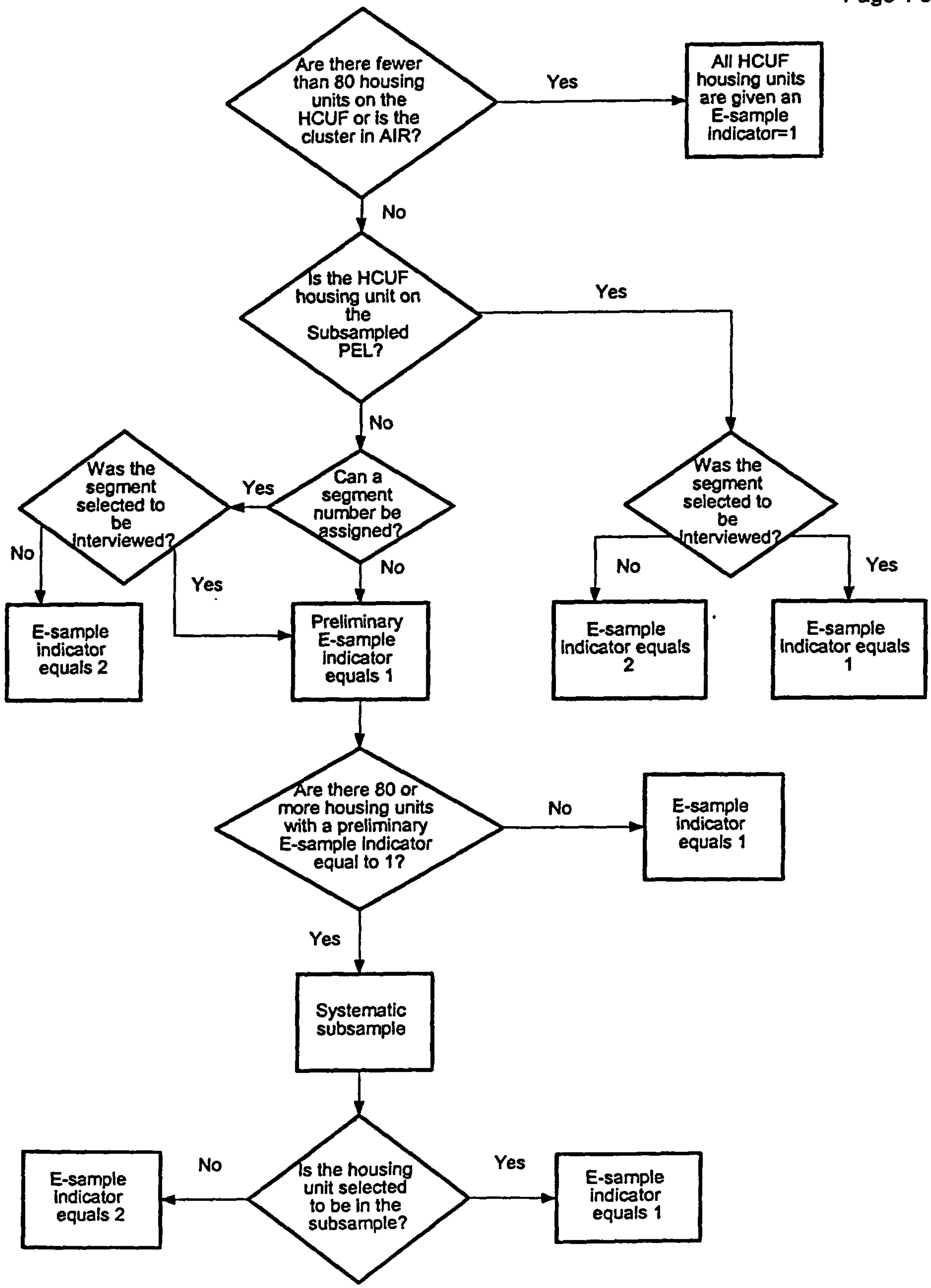
3. DSSD Census 2000 Procedures and Operations Memorandum Series, Chapter S-DT-01, "Accuracy and Coverage Evaluation: The Design Document," January 11, 2000.

cc. DSSD Census 2000 Procedures and Operations Memorandum Series Distribution List
A.C.E. Implementation Team
Statistical Design Team Leaders
Sample Design Team

Description of HU Groups

HU Group Code	Description
1	HCUF HUs in Clusters with < 80 HCUF HUs or in AIR Clusters
2	Corresponding HCUF HUs in Clusters with 80+ HCUF HUs in Segments Selected for the P Sample
3	Corresponding HCUF HUs in 80+ Clusters in Segments Not Selected for the P Sample
4	Non-corresponding HCUF HUs in 80+ Clusters in P-Sample Segments and with No E-Sample Subsampling
5	Non-corresponding HCUF HUs in Clusters with 80+ HCUF HUs in P-Sample Segments and Selected in E-Sample Subsampling
6	Non-corresponding HCUF HUs in 80+ Clusters in P-Sample Segments and Not Selected in E-Sample Subsampling
7	Non-corresponding HCUF HUs in 80+ Clusters in Segments Not Selected for the P Sample
8	HCUF HUs in Special Case Clusters Selected in E-Sample Subsampling
9	HCUF HUs in Special Case Clusters Not Selected in E-Sample Subsampling

Flowchart of the E-sample Identification Process



Layout of the Subsampled Preliminary Enhanced List

```
 Layout Name :  ENHANCED00.LAY                                  Page :    1
 Description :  2000 ENHANCED LIST LAYOUT
Total Length :    360
Date Created :  05-01-2000

                                                               Positions
  #   Field           Field description               length  Beg  -   End
  1.  CNTRLNM         CONTROL NUMBER                    24      1  -    24 CHAR
                         1: 4  LCO
                         5:10  CLUSTER
                        11:12  SEGMENT
                        13:17  MAP SPOT NUMBER
                        18:21  WITHIN MSN ID
                        22:24  ZERO FILL
  2.  LCO             LOCAL CENSUS OFFICE                4     25  -    28 CHAR
                      *****************************
                        Index 1 CLUST thru WMSN
                      *****************************
  3.  CLUST           CLUSTER NUMBER                     6     29  -    34 CHAR
  4.  MSN             ENHANCED IL MAP SPOT NUMBER        5     35  -    39 CHAR
  5.  WMSN            WITHIN MAP SPOT NUMBER ID          4     40  -    43 CHAR
                      *****************************
                        Index 2  CID
                      *****************************
  6.  CID             MAF ID                            12     44  -    55 CHAR
  7.  BLK             1998 BLOCK AND SUFFIX              6     56  -    61 CHAR
  8.  URBNZ           URBANIZATION                      30     62  -    91 CHAR
  9.  HSNUM           HOUSE NUMBER (LJ/BF)              10     92  -   101 CHAR
 10.  SNAME           STREET NAME (LJ/BF)               35    102  -   136 CHAR
 11.  UNIT            UNIT DESIGNATION (LJ/BF)          15    137  -   151 CHAR
 12.  RR              RURAL ROUTE/BOX # (LJ/BF)         25    152  -   176 CHAR
 13.  POBX            PO BOX NUMBER (LJ/BF)             10    177  -   186 CHAR
 14.  CITY            CITY/TOWN NAME                    20    187  -   206 CHAR
 15.  ZIP             ZIP CODE                           5    207  -   211 CHAR
 16.  ZIP4            ZIP + 4                            4    212  -   215 CHAR
 17.  STATE           FIPS STATE ABBREVIATION            2    216  -   217 CHAR
 18.  FIPSCNTY        FIPS COUNTY CODE                   3    218  -   220 CHAR
 19.  FIPST           FIPS STATE CODE                    2    221  -   222 CHAR
 20.  PL              PHYSICAL LOCATION DESCRIPTION     50    223  -   272 CHAR
 21.  PRKNM           TRAILER PARK NAME                 30    273  -   302 CHAR
 22.  HUFIN           MATCH CODE FROM HU MATCHING        2    303  -   304 CHAR
 23.  HUFINID         ID FROM HOUSING UNIT MATCHING     12    305  -   316 CHAR
 24.  TOA             TYPE OF BASIC ADDRESS              1    317  -   317 CHAR
                        1 = ONE FAMILY HOUSE
                        2 = BSA WITH 2 OR MORE HUS
                        3 = MOBILE HOME NOT IN PARK
                        4 = MOBILE HOME IN PARK
                        5 = ONE FAMILY HOME IN
                            SPECIAL PLACE
                        6 = BSA WITH 2 OR MORE HUS
                            IN A SPECIAL PLACE
                        7 = OTHER
 25.  USTAT           UNIT STATUS                        1    318  -   318 CHAR
                        1 = OCCUPIED OR VACANT AND
                            INTENDED FOR OCCUPANCY
                        2 = UNDER CONSTRUCTION
                        3 = FUTURE CONSTRUCTION
                        4 = UNFIT FOR HABITATION
                        5 = BOARDED UP
                        6 = STORAGE OF HOUSEHOLD
```

Layout Name : ENHANCED00.LAY
Description : 2000 ENHANCED LIST LAYOUT
Total Length : 360
Date Created : 05-01-2000

#	Field	Field description	length	Positions Beg	-	End	
		GOODS 7 = VACANT MOBILE HOME SITE 8 = OTHER					
26.	UR	(Not used in 2000)	1	319	-	319	CHAR
27.	QAFLG	QA SAMPLE FLAG 0 = NOT IN QA SAMPLE 1 = IN QA SAMPLE	1	320	-	320	CHAR
28.	ESAMPFLG	E-SAMPLE ELIGIBILITY FLAG	1	321	-	321	CHAR
29.	URFLAG	FLAG INDICATING THAT ADDRESS IS CONSIDERED TO BE URBAN OR RURAL 0 = RURAL 1 = URBAN	1	322	-	322	CHAR
30.	MULTIFLAG	FLAG INDICATING THAT UNIT IS IN A MULTIUNIT OF LESS THAN 20 UNITS 0 = MULTI <20 UNITS 1 = NONMULTI, OR MULTI >= 20	1	323	-	323	CHAR
31.	DSSDSEG	SEGMENT FOR LARGE BLOCK SUBSAM	2	324	-	325	CHAR
32.	FLDSEG	SEGMENT FOR ASSIGNING WORK IN	2	326	-	327	CHAR
33.	INTERVW	AFTER LARGE BLOCK SUBSAMP 0 = OUT OF SAMPLE 1 = IN SAMPLE 8 = SUPP/OUT OF SAMPLE 9 = SUPP/IN SAMPLE	1	328	-	328	CHAR
34.	JIC	JUST IN CASE SPACE ****************************** THESE FIELDS ARE USED FOR LARG BLOCK SUBSAMPLING. ******************************	14	329	-	342	CHAR
35.	TOTCASES	NUMBER OF CASES IN CLUSTER	6	343	-	348	CHAR
36.	ICMCASES	NUMBER OF ICM CASES IN CLUSTER	6	349	-	354	CHAR
37.	CENCASES	NUMBER OF CEN CASES IN CLUSTER	6	355	-	360	CHAR

Layout of the Sample Design File

Variable Description	Name	Places	Source
Census Region	REGION	1	UN
Census Division	DIV	2	UN
State code	STATE	3-4	UN
County code	COUNTY	5-7	UN
Local census office	LCO	8-11	CS
Interim Tract (Pseudo Tract)	ITRACT	12-17	BC
Current Sample Indicator	CSI	19	UO
A.C.E. block cluster number	CLUST	21-25	CS
Check Digit	DIGIT	26	CS
Geography block cluster number	GCLUST	28-32	BC
List/Enumerate Indicator	LEIND	33	BC
Type of Enumeration Area Recode	TEACR	34	CS
Type of Enumeration Area group	TEAG	36	BC
Number of HUs used for sample design	NHU	37-41	BC
Number of MAF HUs	NHUM	43-47	BC
Number of 1990 HUs	NHU90	49-53	BC
Sampling Stratum 1 = Small 2 = Medium 3 = Large 4 = American Indian Reservation	SS	55	UN
American Indian Country Indicator 0 = No American Indian Country 1 = American Indian Reservation/trust land 2 = Tribal Jurisdiction Area/ Alaska Native Village Statistical Area/ Tribal Designated Statistical Area	AICIND	56	BC
Demographic/Tenure Group code	DTCODE	57-58	UN
Demographic/Tenure Group label	DTLABEL	59-60	UN
Estimated Urbanicity of block cluster 1 = Urban Area with population ≥250,000 2 = Other Urban Area 3 = Non-Urban Area	ECLUSURB	62	UN
Size Category 1=Small (0-2 hus) 2=Medium (3-79 hus) 3=Large (80+ hus)	SIZCAT	63	UN
Additional space		64-91	

Variable Description	Name	Places	Source
First step index number	INDEX1	92-99	CS
Listing sample selection indicator	BC1	101	CS
1 = Selected			
Random Start for listing sample selection	RS1	103-113	UN
Take-every for listing sample selection	TE1	115-125	UN
Second step listing sample selection indicator	BC2	127	CS
0 = Not Selected			
1 = Selected			
Random Start for the second step of the listing sampling	RS2	129-139	CS
Take-every for the second step of the listing sampling	TE2	141-151	CS
Unbiased weight after block cluster sampling	WEIGHTBC	153-164	CS
Additional space		165-175	
Preliminary Number of HUs on the Independent List	NHUILP	176-180	AR
Number of Housing Units On the January 2000 DMAF	NHUDMAF	182-186	AR
Demographic Code	DEMCODE	188	AR
1 = Minority			
2 = Non-Minority			
3 = Puerto-Rico			
Consistency Code	CONCODE	189	AR
1 = Low Inconsistent (IL significantly smaller than DMAF)			
2 = Consistent			
3 = High Inconsistent ((IL significantly larger than DMAF)			
A.C.E. Reduction Stratum	ARST	190-191	AR
A.C.E. Reduction Indicator	ACERED	193	AR
0 = Not Selected			
1 = Selected			
Random Start for A.C.E. Reduction	RSAR	195-205	AR
Take-every for A.C.E. Reduction	TEAR	207-217	AR
Unbiased weight after A.C.E. reduction	WEIGHTAR	219-230	AR
Collapsing Flag	COLFLAG	232	AR
A.C.E. Reduction Index Number	INDEXR	234-241	AR
Number of Housing Units On the December 1999 DMAF (Initial)	NHUDMAFI	243-247	AR
Additional space		248-300	
Number of HUs on the Independent List	NHUIL	301-305	SB
Small Block Cluster Subsampling Stratum	SBCSS	306-307	SB
Small Block Subsampling Indicator	SB	308	SB
0 = Not Selected			
1 = Selected			
Random Start for Small Block subsampling	RSSB	310-320	SB
Initial take-every for Small Block subsampling	ITESB	322-332	SB
Unbiased weight for A.C.E. cluster	WEIGHTC	334-345	SB
Larger of the DMAF and IL HU count	LARGERHU	347-351	SB
Final take-every for Small Block subsampling	FTESB	352-362	SB
Additional space		363-370	

Variable Description	Name	Places	Source
Relisted Block Cluster Flag 0 = Not Relisted, 1 = Relisted	RELIST	371	LB
Number of total hus in block cluster	NHUEL	373-377	LB
Number of A.C.E. hus in cluster	NHUELA	379-383	LB
Number of supplemental hus in cluster	NHUELN	385-389	LB
Large Block Cluster EL subsampling code 1 = NHUELI< 80 hus, 2 = NHUELI ≥ 80 hus	ELLBSUB	391	LB
Random Start for Large Block subsampling	RSLB	393-403	LB
Take-every for Large Block subsampling	TELB	405-415	LB
Number of segments in block cluster	NSEG	417-418	LB
Number of segments selected in block cluster	NSEGSAM	420-421	LB
Day of Arrival	DAY	423-424	LB
Final Cluster Order Number	CON	431-434	LB
Number of total hus for interview in block cluster	NINT	436-440	LB
Unbiased weight for P-sample HUs	WEIGHTP	442-453	LB
Number of Assignments in block cluster	NA	455-456	LB
Final Sampling Strata	FSS	458-464	LB
Additional space		465-490	
Number of CUF HUs in block cluster with an ESPS code of 1	NHUCUF1	491-495	ES
Number of CUF HUs in block cluster with an ESPS code of 2	NHUCUF2	497-501	ES
Number of CUF HUs in block cluster	NHUCUF	503-507	ES
Number of CUF HUs in selected segments with an ESPS code of 1	NHUCUFS1	509-513	ES
Number of CUF HUs in selected segments with an ESPS code of 2	NHUCUFS2	515-519	ES
Number of CUF HUs in selected segments of a block cluster	NHUCUFS	521-525	ES
E-Sample Identification cluster category 1 = NHUCUF < 80 2 = NHUCUF ≥ 80 and NHUCUFS < 80 3 = NHUCUF ≥ 80 and NHUCUFS ≥ 80 4 = NHUCUF ≥ 80 and RELIST = 1 5 = NHUCUF ≥ 80 and List/Enumerate 6 = NHUCUF ≥ 80 and only UI/CI SPEL HUs 7 = NHUCUF ≥ 80 and zero SPEL HUs	EICC	527	ES
Random Start for E-sample subsampling	RSES	529-539	ES
Take-every for E-sample subsampling	TEES	541-551	ES
Number of E-sample HUs in block cluster with an ESPS code of 1	NHUES1	553-557	ES
Number of E-sample HUs in block cluster with an ESPS code of 2	NHUES2	559-563	ES
Number of E-sample HUs in block cluster	NHUES	565-569	ES
Unbiased weight for E-sample HUs with an ESPS code of 1	WEIGHTE1	571-582	ES
Unbiased weight for E-sample HUs with an ESPS code of 2	WEIGHTE2	584-595	ES

Variable Description	Name	Places	Source
Number of confirmed A.C.E. housing units not found in the census	CURCI	676-680	TES
Number of unconfirmed A.C.E. housing units not found in the census	CURUI	682-686	TES
Number of census housing units geocoded to the wrong census block	CURGE	688-692	TES
Targeted extended search selection type	TESSELECT	694	TES
Targeted extended search selection flag	TESFLAG	696	TES
Random Start for the targeted extended search	RSTES	698–709	TES
Take-every for the targeted extended search	TETES	710–721	TES
Targeted Extended Search Index Number	TESN	722-727	TES
Additional Space		728-750	

Source Codes

AR: A.C.E. Reduction
BC: Block Clustering
CS: Block Cluster Sampling
ES: E-sample Identification
LB: Large Block Subsampling
SB: Small Block Subsampling
UN: Universe File Creation
UO: Updated for each operation
TES: Targeted Extended Search

E-Sample Codes and Probability of Selection Outcomes

Code	Number of HCUF units in the A.C.E. Block Cluster									
	< 80		80+							
			CUF HU Corresponds with HU on the SPEL		CUF HU Does Not Correspond with HU on the SPEL					
	Segment In Sample	Segment Not In Sample	Segment In Sample	Segment Not In Sample	Can Assign to Segment				Cannot Assign to Segment (special case cluster)	
					Segment In Sample			Segment Not In Sample		
					<80 Non-corr.	80+ Non-corr.				
						Samp	Not Samp		Samp	Not Samp
HU Group	1	1	2	3	4	5	6	7	8	9
Preliminary E-Sample Indicator	0	0	0	0	1	1	1	0	1	1
E-Sample Indicator	1	1	1	2	1	1	2	2	1	2
E-Sample Prob. Code	1	1	1	1	2	2	2	2	2	2
Weight	WEIGHTC	WEIGHTC	WEIGHTP	NA	WEIGHTP	WEIGHTE2	NA	NA	WEIGHTE2	NA

See the following page for notes on this table.

Notes:

1. In HU Groups 5 and 8, "Samp" means the HCUF HU was selected during E-sample subsampling. Likewise "Not Samp" in HU Groups 6 and 9 means the HCUF HU was not selected.

2. WEIGHTE2 in HU Group 5 is TE1×TE2×TEAR×FTESB×TELB×TEES. WEIGHTE2 in HU Group 8 is TE1×TE2×TEAR×FTESB×TEES.

3. WEIGHTC is the A.C.E. block cluster weight after small block cluster subsampling.
WEIGHTP is the A.C.E. block cluster weight after large block cluster subsampling.
TEES is the take-every used for E-sample subsampling.
TE1, TE2, TEAR, FTESB, and TELB are take-everys from previous sampling operations and are obtained from the Sample Design File.

4. Possible combinations of weights within a block cluster are
   - All HUs have WEIGHTC
   - All HUs have WEIGHTP
   - All HUs have WEIGHTP×TEES
   - All HUs have WEIGHTC×TEES
   - Corresponding HUs have WEIGHTE1 = WEIGHTP and non-corresponding HUs have WEIGHTE2 = TE1×TE2×TEAR×FTESB×TELB×TEES
   - Corresponding HUs have WEIGHTE1 = WEIGHTP and non-corresponding HUs have WEIGHTE2 = TE1×TE2×TEAR×FTESB×TEES

5. If E-Sample Probability Code =1 use the WEIGHTE1 variable on the Sample Design File to apply the appropriate weight
=2 use the WEIGHTE2 variable on the Sample Design File to apply the appropriate weight.

6. In HU Groups 3, 6, 7, and 9, "NA" means an E-sample weight is not applicable because the HU Group is not in the E sample.

Example of Segment Identifier Assignment

The table on the next page contains an example of a block cluster with 88 HCUF HUs that correspond to HUs on the SPEL. This is a large cluster, but only a few of the HUs are shown for illustrative purposes.

The HCUF HUs that correspond with HUs on the SPEL have a segment identifier. The non-corresponding HCUF HUs are sorted into their proper order using the house number and street name. They are assigned the segment identifier of the previous corresponding HCUF HU. The segment(s) selected for the A.C.E. interview sample is also the segment(s) selected for the E sample to achieve an overlapping E sample and P sample. The corresponding HCUF HUs in selected segments are assigned an E-sample indicator equal to one. The non-corresponding HCUF HUs in selected segments are assigned a preliminary E-sample indicator equal to one and may be subject to subsampling if there are 80 or more of them. The HCUF HUs in non-selected segments are assigned an E-sample indicator equal to two.

Segment Identifier Example in a Block Cluster with City-Style Addresses

SPEL A.C.E. Map Spot Number	Segment Identifier from SPEL	Segment Identifier Assigned during E-sample Identification	Address
11	BA		101 1st St
12	BA		101A 1st St
13	CA		103 1st St
		CA	105 1st St
4	AA		309 Maple Ln
5	AA		311 Maple Ln
		AA	311 Maple Ln Basement
7	BA		315 Maple Ln
1	AA		104 Peach Ct
		AA	110 Peach Ct
		AA	112 Peach Ct
		AA	116 Peach Ct
		AA	120 Peach Ct
2	AA		702 Vermont Ave
		AA	704 Vermont Ave
3	AA		704 Vermont Ave rear
		AA	704 Vermont Ave Upper
⋮	⋮	⋮	⋮

ALLOCATION OF THE ICM SAMPLE TO THE STATES FOR CENSUS 2000

Eric Schindler, Bureau of the Census
Bureau of the Census, Washington, DC 20233

KEYWORDS: Dual System Estimation, Reapportionment, Jackknife

ABSTRACT: The introduction of Integrated Coverage Measurement (ICM) for Census 2000 requires 51 state estimates based only on data from each state. The goal is to allocate the available sample of 750,000 housing units so as to achieve coefficients of variation for the Dual System Estimates of 0.5% in all states and standard errors of about 60,000 in the larger states. Data from the 1990 Post Enumeration Survey are restratified and dual system estimates with Jackknife variances are calculated. The need for good data quality in both the initial phase and the ICM phase and the effect on Congressional reapportionment are also discussed.

I. Introduction

Census 2000, as currently planned, will integrate the results of a large coverage survey into the final census estimates at all levels of geography. This paper describes the applied research used to determine an appropriate allocation of the Integrated Coverage Measurement (ICM) sample to the states for Census 2000. For more information on Census 2000 and the design of the ICM program, see Hogan and Waite (1998) or Griffin and Vacca (1998). The following basic facts are considered by the design:

- The total ICM sample size will be about 750,000 housing units. This size was determined by the Census Bureau's ability to implement and control the ICM sample and by statistical requirements. Rough preliminary estimates indicated that this sample size might be enough to produce coefficients of variation of 0.50% in each state. Block clusters averaging about 30 housing units will continue to be the primary sampling unit. The total ICM sample will have about 25,000 block clusters. Data from an independent second enumeration of the ICM block clusters will be compared to the Initial Phase estimate using Dual System Estimation. In comparison, the 1990 Post Enumeration Survey (PES) was only about one-fifth as large.
- A Supreme Court ruling in March 1996 and others have expressed concern about the PES state level total population estimates based on data from several states. The official population of each state and the District of Columbia released on December 31, 2000 will be estimated directly from the data from within the state.
- The primary goal of ICM is to improve the accuracy of the Congressional reapportionment process. In statistical terms, the expected value of most state population estimates should be closer to the true value with ICM than with a traditional census. Without the ICM, the wrong states in terms of their true populations may be competing for the last few seats in the House of Representatives. With ICM, the right states are more likely to be in the competition.
- The primary goal of ICM allocation is to optimize the precision of the apportionment process. In statistical terms, ICM allocation attempts to make the state population estimates close to their expected values[1]. Optimizing the precision of the apportionment process for Census 2000 would require decreasing the standard errors of those four to six states competing for the last three or four seats in the House of Representatives as much as possible and virtually ignoring the other states. However, precensal estimates will not be accurate enough to identify these last few states.
- Since census data are also used for redistricting, for allocation of federal and state funds, for planning purposes, etc., reasonable precision is also required for those states whose apportionment is certain and for synthetic estimates for substate areas and population subgroups.

Section II describes the research leading to the

[1] The 1990 reapportionment based on census counts was more precise (closer to its expected value) but less accurate (expected value missed the true value) than the apportionment process will be with ICM in Census 2000. Initial Phase estimates will be close to their expected values which may be far from the true population. ICM state estimates will miss their expected values, which will be closer to the true values, by more than the Initial Phase estimates miss their expected values. However, the ICM estimates will generally miss the "true" values by less than the Initial Phase estimates miss the "true" values.

The author is a mathematical statistician in the Decennial Statistical Studies Division of the US Census Bureau. The author wishes to thank John Thompson, Henry Woltman, Richard Griffin, and Alfredo Navarro for their guidance and insight over the last several years.

recommended state ICM sample sizes using data from the 1990 Post Enumeration Survey (PES). Section III discusses the possible effect of changes in data quality from 1990 to 2000. Section IV discusses the effect of ICM sampling errors on congressional reappportionment. Section V provides a brief summary.

II. Methodology

Step 1: Redefine Sampling Strata:

For the 1990 PES 112 sampling strata were defined based on the Census division, degree of urbanization, minority population, and tenure. Some of these sampling strata had very small sample sizes. For this work the 1990 PES sampling strata were collapsed to 39 sampling strata. In addition to a national stratum for American Indians living on reservations, each of nine census divisions has zero, one, or two minority redefined strata (total 13, none in New England or the North Central division), and two or three non-minority redefined strata (total 25). Each state has PES block clusters in from two to six of the redefined sampling strata. There are 186 sampling stratum/state substrata for non-American Indian Reservations and 14 for the American Indian Reservations.

Step 2: Remove Outlier Block Clusters:

Thirty-nine block clusters which contribute heavily to the error were identified and removed. These clusters generally have high sampling weights and accounted for a large portion of the undercount or overcount in the sampling stratum/state cell. Outliers of the magnitude found in 1990 could as much as double the standard errors of the affected states. Several proposed design changes will help to control the effect of outliers in Census 2000:

- In 1990 large block clusters (over 80 housing units) were subsampled before PES collection. The subsampling resulted in high weights. Increasing the number of large block clusters in the 2000 ICM will reduce their initial weights. Subsequent subsampling will increase the weights back to a normal level.
- In 1990 only a very small sample of small block clusters (0-2 housing units) was selected. During PES collection some of these block clusters were found to be much larger, giving high weights to a large number of housing units. In 2000 a two stage sample for very small block clusters will control the weights of those block clusters which are found to be much larger than expected.

These approaches involve additional costs and will not succeed completely in eliminating the effects of outlier block clusters.

Step 3: Adjust Weights to Match State Totals:

The PES E-Sample consists roughly of the 1990 census reports of those persons in the PES sample block clusters. For each block cluster, weighted estimates of the E-Sample are calculated. For many states the weighted E-Sample is not a good estimate of the 1990 census count. For each state the sum of the state's weighted E-Sample estimate is ratio adjusted to the 1990 census count. The weighted estimates of erroneous enumerations (persons in the census who should not have been counted), P-Sample (persons enumerated in the second interview in the PES block clusters), and omissions (persons in the P-Sample who could not be matched back to the census) are multiplied by the same ratio.

Step 4: Make State and Stratum Estimates:

For each of the 39 sampling strata, dual system estimates are calculated by:

$$DSE_{Div,k} = C_{Div,k} \frac{E_{Div,k} - EE_{Div,k}}{E_{Div,k}} \frac{P_{Div,k}}{M_{Div,k}}$$

where:

$C_{Div,k}$	is the census count in stratum k or in this case the weighted E-sample,
$E_{Div,k}$	is the E-sample estimate in stratum k for the census division,
$EE_{Div,k}$	is the estimated number of erroneous enumerations in stratum k for the census division,
$P_{Div,k}$	is the P-sample estimate in stratum k for the census division, and
$M_{Div,k}$	is the estimated number of P-sample persons who match to the E-sample in stratum k for the census division.

A Jackknife procedure dropping one block cluster at a time from each census division's PES sample without reweighting[2] is used to estimate standard errors, $SE_{Div,k,n_{k,Div}}$, and variances, $VAR_{Div,k,n_{k,Div}}$, for the E-sample person sample sizes, $n_{k,Div}$, for the DSE in stratum k for division Div.

Since the finite population correction factors are negligible, for the same sample size, $n_{k,Div}$, the CV of state i for stratum k is the same as the CV for the division.

---

[2] The DSE and its population variance can also be estimated by Taylor Series expansion from the erroneous enumeration and omission rates. The results are consistent with those of the approach used here. This more direct approach is preferred because it is simpler and it is consistent with the 1990 and 2000 variance estimation methods. Other options considered included equal allocations to all states and various combinations of the alternatives.

Therefore: $SE_{i,k,n_{kDiv}} = SE_{Div,k,n_{kDiv}} \dfrac{E_{i,k}}{E_{Div,k}}$.

Step 5: Determine Block Cluster Sample Sizes

We assume n_i^0 = 10,000 E-Sample persons.

Allocating these persons proportionally to the states's E-sample population in the redefined strata we have:

$$n_{i,k}^0 = 10000 \frac{E_{i,k}}{\sum_{k'} E_{i,k'}}.$$

The standard errors for these stratum sample sizes are: $SE_{i,k,n_{i,k}^0} = SE_{i,k,n_{kDiv}} \sqrt{\dfrac{n_{kDiv}}{n_{i,k}^0}}$ for each stratum and

$$SE_{i,n_i^0} = \sqrt{\sum_k SE^2_{i,k,n_{i,k}^0}}$$ for the state total.

The next step is to convert the $n_{i,k}^0$ to $b_{i,k}^0$, the number of block clusters in state i stratum k, using the observed average block cluster E-Sample person size for stratum k within Census division.

If SE_{i,b_i^0} is the standard error for the block cluster sample size b_i^0 corresponding to the 10,000 E-Sample persons, the sample size, in block clusters, required to obtain a desired standard error, SE_i is: $b_i = b_i^0 \dfrac{SE^2_{i,b_i^0}}{SE_i^2}$.

An allocation of 18,873 block clusters is required to achieve the desired coefficients of variation (CV) of 0.5% in all states. These allocations are shown in column 5 of Table 1. [3]

Step 6: Assure Minimum Sample Size

Thirteen states have ICM samples less than 300 block clusters from step 5. These states are concentrated in several divisions with relatively low estimated population variance. Since the estimated population variances, which are subject to high variance, may not be as low in 2000 as in 1990, the samples sizes for these states are increased to 300 block clusters to be more in line with the remaining states. These increases require about 1200 block clusters, increasing the total allocated so far to about 20000. The results are shown in columns 6 and 7 of Table 1.

Step 7: Reduce Expected Standard Errors for States with Populations over 10,000,000

Reserving 350 block clusters for American Indian Reservations, about 4600 block clusters remain to be assigned. These are assigned to the largest states proportionately to the squares of their 1990 census counts[4]. This reduces the estimated standard errors of the largest states from 0.50% of their population to 55672. These decreases are particularly substantial in the largest states: California, New York, and Texas. The sample size for Ohio was increased from 260 to 358, executing steps 6 and 7 simultaneously. The results are shown in columns 8 to 10 of Table 1. Columns 11 and 12 show the number of persons and occupied housing units which can be expected in each state.

AMERICAN INDIAN RESERVATIONS

In 1990 the largest reservations, spread across fourteen states, were covered by a sample of 43 block clusters. American Indians living on reservations or other tribal lands have special legal status. In 1990 variances were high for this hard to count population of about 800,000 people with about a 10% undercount rate. 350 block clusters, about as many as the states with fewer than 10,000,000 residents, 1.4% of the sample, were set aside for this 0.3% of the population.

III. ICM Quality Concerns

The ICM sample sizes calculated above are designed to yield errors of 0.5% or 56,000, whichever is smaller in all states. Table 1 shows that California would require 361 block clusters to achieve a CV of 0.5%. Estimates made for the state of California show a CV of about 0.45% for its 381 1990 PES block clusters. On the other hand, the CVs calculated for the 1995 and 1996 tests were considerably higher than the design estimates. The DSE is roughly the

---

[3] The allocations for states in the same Census Division are correlated because the same population variance estimates are being used. The differing proportions of the population in each sampling stratum account for the small differences between states. It is possible to repeat this procedure entirely within each state. This eliminates the synthetic estimation from the Census Divisions to the states and the correlation of the allocations. Unfortunately, most stratum/state cells do not have sufficient sample to obtain reliable estimates.

[4] The use of projected 2000 populations was considered, but the estimated allocations for several states seemed inappropriate.

Initial Phase estimate times the rate of Initial Phase persons who are correctly enumerated times the inverse of the rate of P-Sample persons who could be matched back to census reports:

$DSE = IP \times R_{CE} \times 1 / R_{MATCH}$ where both rates are close to 1. There is comparatively little variance in IP, so (assuming equal design effects and even with some correlation) the variance is proportional to the sum of two PQ type variances:

$$VAR_{DSE} \approx \frac{R_{CE}\times(1-R_{CE})}{n} + \frac{R_{MATCH}\times(1-R_{MATCH})}{n}$$

where n is the ICM sample size[5].

There are several operational changes from 1990 in the design for Census 2000 which may decrease either the correct enumeration rate and/or the match rate. A 3% decrease in both the correct enumeration rate and the match rate from 97% to 94% would not change the estimate much, but it could double the estimated variance, multiplying the estimated CV by about 1.4. These changes include:

- The easy availability of Be Counted Forms could increase the number of erroneous enumerations, decreasing the correct enumeration rate.
- The use of a five person form instead of a seven person form could increase the number of persons, especially children and nonrelatives, missed in the initial phase, decreasing the match rate.
- The tight schedule and decreased public cooperation could increase the number of whole household imputations in the initial phase, decreasing the match rate. The rate was about 1% in 1990 but about 8% in the 1996 test in Chicago.
- Not performing a surrounding block search for additional matches or performing a limited surrounding block search could decrease the match rate.

There are few counterbalancing changes to improve the data quality[6]. Therefore, it should be expected that the calculated standard errors for Census 2000 may be somewhat higher than those predicted by the design.

[5]There is a third ratio in the DSE formula which adjusts for whole person imputations in the Initial Phase. This term adds little to the variance, but it corrects for census persons who cannot be matched to by P-Sample persons because their census data is imputed.

[6]The Be Counted Forms should decrease the number of nonmatches and decreased weight variation should make the sample more efficient.

IV. Effect on Reapportionment

The 435 seats in the House of Representatives are reapportioned to the states using the Hill Algorithm which works as follows:

- Assign each state one seat.
- For each state calculate: $R_i = POP_i / \sqrt{N_i\times(N_i+1)}$

  where POP_i is the population of state i being used in the apportionment, and N_i is the number of seats already assigned to state i.
- Assign the next seat to the state with the largest value of R_i
- Calculate new R_is and repeat the process until all 435 seats are assigned.

Using the 1990 PES instead of the 1990 census counts would have given one more seat to California at the expense of Wisconsin. For the two apportionments the 435th seat was assigned as follows:

- 1990 census count: Washington
  Massachusetts was next and would have needed 12605 more inhabitants to take the last seat instead of Washington.
- 1990 PES estimate: Pennsylvania
  Wisconsin was next and would have required 12573 more inhabitants to take the last seat. Montana was fourth in line but would have required only 3919 more inhabitants to take the last seat.

The 1990 PES estimate for state i can be viewed as a random draw from the normal distribution about the true value. That is: PES_i is selected from $N(T, SE_{PES,i})$. Since we know PES_i, we can reverse the situation and obtain 100 possible target values of the truth for each state, i, by selecting T_{ij}, j=1,100 from $N(PES_i, SE_{PES,i})$. For each target estimate T_{ij}, we can obtain 100 estimates of possible values that the ICM would produce, ICM_{ijk}, by sampling from $N(T_{ij}, SE_{ICM,i})$. Thus, it is possible to compare the apportionment from the 1990 Census to 100 1990 targets and the apportionments from 10000 ICM estimates to the same 100 1990 targets. The results are shown in table 2.

- Using either census counts or ICM, there is only a small probability that the apportionment process will assign all 435 seats to the correct states.
- Over the 10000 simulations of ICM estimates, the 1990 census and 2000 ICM apportionments had the same number of errors compared to the target "true" apportionments 3738 times. The 1990 census apportionment had fewer errors 1982 times. The 2000 ICM apportionments had fewer errors 4280 times.
- Over the 10000 simulations there were 42032 instances where a state had a difference between its 1990 census apportionment and its 2000 ICM apportionment. In these instances, the 1990 census apportionment matched the target apportionment

18270 (43.47%) times. The 2000 ICM apportionment matched the target apportionment 23762 (56.53%) times.

- Using the 1990 PES estimates or the ICM estimates, the states with the most variation in the target apportionments; that is, the states which may deserve one more or one less seat, are bunched around the 435th selection for both the target and the ICM apportionments. On the other hand, even though there is only one difference between the 1990 census and the 1990 PES apportionments, the states with variation in their target apportionments are not the states clustered around the 435th selection in the 1990 census data apportionment.

Table 2: Number of Seats Shifted Compared to Target Apportionments over 100 Simulations for 1990 Census or 10000 Simulations for 2000 ICM

	1990 Census	2000 ICM
Census or ICM apportionment equals target apportionment	3	1285
One seat shifted by census or ICM apportionment compared to target apportionment	41	5015
Two seats shifted	51	3104
Three seats shifted	5	554
Four seats shifted	0	41
Five seats shifted	0	1
Average number shifted	1.58	1.31

V. Summary

- For the allocation proposed based on the underlying population variances, it is estimated that a total ICM sample size of 24,650 block clusters (if allocated appropriately and assuming data quality equivalent to 1990) is sufficient to (1) achieve coefficients of variation of 0.50% in states with populations less than 10,000,000, (2) allocate each state at least 300 block clusters, and (3) achieve standard errors of 55672 for states with a population over 10,000,000.
- The expected CVs or SEs calculated above are just that: EXPECTED. The increase in population since 1990 will increase the standard errors of the largest states from 55672 to about 60000. Estimates show that the CVs of the estimated CVs or SEs are about 20%. That means that, even if the average state CV is 0.50%, about 16% of the states (8 states) will likely have CVs or SEs at least 20% larger than the expected values or above 0.60% or 72,000; and about 2% (1 state) will likely have a CV or SE 40% larger than the expected value or above 0.70% or 84,000. Any decrease in data quality compared to 1990 may further increase the CVs or SEs in 2000.
- The proposed ICM sample sizes in 2000 will be sufficient to assure that the correct states are in the competition for the last few seats in the House of Representatives, but they will not be sufficient to assure that all 435 seats are apportioned perfectly. The apportionment process is very sensitive to minor population variations and no affordable ICM sample size can reduce the standard errors enough to assure perfect apportionment. However, a traditional census count would virtually insure that the apportionment would be incorrect.
- It is necessary to explain this technical decision to non-technical audiences. This option is relatively simple. Since it is similar to the variance estimation methods for 1990 and 2000, it should already be familiar to many of the involved parties.

Issues which have not been investigated are:

- Only the variance from the ICM sample has been considered. Variance from sampling for nonresponse follow-up and housing units returned as vacant by the post office, will be small at the state level and will have no significant effect on the estimates. A third source, variance due to imputation of missing data, could be more substantial.
- For the allocation of the ICM sample within each state, sampling strata and estimation poststrata will be developed to permit adequate estimates for race, Hispanic origin, age, sex, tenure, and geographic subpopulations. Oversampling small but visible subpopulations could increase the state level errors.

References:

Hogan, H. and Waite, J. (1998) "Statistical Methodologies for Census 2000 - Decisions Issues, and Preliminary Results," *Proceedings of the Survey Research Methods Section, American Statistical Association*, Alexandria, VA, American Statistical Association, to appear.

Griffin, R. and Vacca, E.A. (1998) "Estimation in the Census 2000 Dress Rehearsal," *Proceedings of the Survey Research Methods Section, American Statistical Association*, Alexandria, VA, American Statistical Association, to appear.

TABLE 1: ICM Sample Sizes for CV/SE Combinations For Three Variance Alternatives

CEN DIV	State		1990 Census	Estimated DSE	0.5% CV	300 ClusterMin		SE for States>10000000			Persons	Occ HUs
					18873	20042	CV	24650	CV	SE	1835172	690062
(1)	(2)		(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
1	CT	9	3287116	3343185	377	377	0.500%	377	0.500%	16716	27942	10406
New	ME	23	1227928	1246851	309	309	0.500%	309	0.500%	6234	21676	8365
England	MA	25	6016425	6118807	375	375	0.500%	375	0.500%	30594	27790	10359
	NH	33	1109252	1126308	307	307	0.500%	307	0.500%	5632	21544	8320
	RI	44	1003464	1020492	373	373	0.500%	373	0.500%	5102	27607	10302
	VT	50	562758	571129	285	300	0.487%	300	0.487%	2782	20687	8074
2	NJ	34	7730188	7863038	461	461	0.500%	461	0.500%	39315	32003	12067
Middle	NY	36	17990454	18228030	470	470	0.500%	1261	0.305%	55672	96137	35936
Atlantic	PA	42	11881642	12089483	496	496	0.500%	585	0.461%	55672	41216	15514
3	DE	10	666168	687353	413	413	0.500%	413	0.500%	3437	29380	11062
South	DC	11	606900	640519	384	384	0.500%	384	0.500%	3203	28914	10993
Atlantic	FL	12	12937926	13330671	363	363	0.500%	520	0.418%	55672	38813	14775
	GA	13	6478216	6670979	399	399	0.500%	399	0.500%	33355	28240	10852
	MD	24	4781468	4933145	368	368	0.500%	368	0.500%	24666	27528	10384
	NC	37	6628637	6816377	400	400	0.500%	400	0.500%	34082	28330	10827
	SC	45	3486703	3601446	422	422	0.500%	422	0.500%	18007	29587	11230
	VA	51	6187358	6355420	371	371	0.500%	371	0.500%	31777	27033	10368
	WV	54	1793477	1834188	425	425	0.500%	425	0.500%	9171	28548	11113
4	AL	1	4040587	4132465	417	417	0.500%	417	0.500%	20662	26712	10207
East	KY	21	3685296	3725204	447	447	0.500%	447	0.500%	18626	28588	10939
South	MS	28	2573216	2635900	402	402	0.500%	402	0.500%	13179	26432	9698
Central	TN	47	4877185	4979805	433	433	0.500%	433	0.500%	24899	27716	10786
5	AR	5	2350725	2396511	494	494	0.500%	494	0.500%	11983	35395	13204
West	LA	22	4219973	4309683	595	595	0.500%	595	0.500%	21548	43282	15929
South	OK	40	3145585	3208831	426	426	0.500%	426	0.500%	15891	30590	11598
Central	TX	48	16986510	17418396	795	795	0.500%	1945	0.320%	55672	143215	52277
6	IL	17	11430602	11589356	351	351	0.500%	380	0.480%	55672	29531	11040
East	IN	18	5544159	5568146	276	300	0.479%	300	0.479%	26687	21839	8362
North	MI	26	9295297	9365360	317	317	0.500%	317	0.500%	46514	23746	8997
Central	OH	39	10847114	10917940	260	358	0.427%	358	0.427%	46574	26135	10020
	WI	55	4891769	4915103	288	300	0.490%	300	0.490%	23477	21772	8329
7	IA	19	2776755	2806367	151	300	0.355%	300	0.355%	9959	21029	8206
West	KS	20	2477574	2510872	157	300	0.361%	300	0.361%	9077	21569	8301
North	MN	27	4375099	4434536	145	300	0.348%	300	0.348%	15405	21759	8330
Central	MO	29	5117073	5185737	158	300	0.363%	300	0.363%	18841	21543	8296
	NE	31	1578385	1600796	175	300	0.382%	300	0.382%	6013	21216	8244
	ND	38	638800	647888	149	300	0.352%	300	0.352%	2175	21007	8202
	SD	46	696004	705880	162	300	0.368%	300	0.368%	2440	20612	8137
8	AZ	4	3665228	3783790	492	492	0.500%	492	0.500%	18384	36175	13874
Mountain	CO	8	3294394	3390812	479	479	0.500%	479	0.500%	16954	35754	13589
	ID	16	1006749	1050778	412	412	0.500%	412	0.500%	4763	33979	12612
	MT	30	799065	833975	420	420	0.500%	420	0.500%	3762	34342	12841
	NV	32	1201833	1239255	468	468	0.500%	468	0.500%	6196	35765	13650
	NM	35	1515069	1570331	481	481	0.500%	481	0.500%	7155	35291	13267
	UT	49	1722850	1773525	478	478	0.500%	478	0.500%	8617	35836	13631
	WY	56	453588	470907	418	418	0.500%	418	0.500%	2355	34251	12783
9	AK	2	550043	567494	334	334	0.500%	334	0.500%	2837	27291	10407
West	CA	6	29760021	30769103	361	361	0.500%	2753	0.181%	55672	232642	84161
	HI	15	1108229	1142269	283	300	0.485%	300	0.485%	5545	24272	9308
	OR	41	2842321	2927930	320	320	0.500%	320	0.500%	14640	25993	9830
	WA	53	4866692	5007784	332	332	0.500%	332	0.500%	24883	26918	10060

SAMPLE DESIGN FOR THE CENSUS 2000 ACCURACY AND COVERAGE EVALUATION

Randal ZuWallack, Matthew Salganik, and Vincent Thomas Mule, Jr., U.S. Census Bureau
Randal ZuWallack, U.S. Census Bureau, Rm 2501, Bldg 2, Washington DC 20233

Keywords: Accuracy and Coverage Evaluation, Census 2000, Stratified Systematic Sampling

Introduction

Every ten years the Census Bureau attempts to enumerate every person living in the United States. Although a complete count is desired, past experience indicates it is virtually unattainable. According to past census evaluations using demographic analysis, the undercount has ranged from 2.8 million in 1980 to 7.5 million in 1940 (Bureau of the Census, 1997). Beginning with the 1950 census, the Census Bureau began conducting post-enumeration evaluations to estimate census coverage. These evaluations took a case by case matching approach to identify people who were missed and those who were counted. More recent evaluations of this type include the 1980 Post-Enumeration Program (PEP) and the 1990 Post-Enumeration Survey (PES). For the PEP, information based primarily on the Current Population Survey was used to estimate people not counted in the census enumeration (Fay, 1988). A second part of the PEP involved selecting a sample of census records to estimate the number of erroneous census enumerations. Improvements were introduced for the 1990 PES. Rather than using information that was not specifically designed for measuring census omissions, a survey was designed with this sole purpose in mind. As was done in 1980, a sample was also selected for estimating erroneous census enumerations.

In the tradition of improving census evaluations, the Census Bureau is conducting the Accuracy and Coverage Evaluation (A.C.E.) following the Census 2000 enumeration. Similar to the PES, the A.C.E. checks the quality of the census in two ways. One is by comparing data from the census to data collected from an independent sample of housing units to estimate the number of people missed. The other is by selecting a sample of census records to estimate the number of erroneous census enumerations. This information is combined to determine dual system estimates of the total population and many demographic groups, which is then compared to the census results to estimate coverage rates. This paper discusses all phases of the A.C.E. sample design, how the design was effected by the recent Supreme Court decision on sampling for the Census (Department of Commerce v. United States House of Representatives, 1997), and changes made to the design based on an evaluation of the Census 2000 Dress Rehearsal design.

P Sample and E Sample

Because there are two types of coverage errors, missed people and erroneous inclusions, two samples are selected to evaluate census coverage --the population sample (P Sample) and the enumeration sample (E Sample). The P Sample consists of the people living in the housing units designated for A.C.E. interviews. These units are randomly selected from an address list which is compiled independently of the census list for a sample of geographic areas. The list is referred to as the Independent List. The P-sample people are matched back to the census to determine if they were counted or missed. The E Sample consists of people living in a sample of housing units enumerated in the census. The E-sample people are checked to determine whether they were correctly counted in the census, or whether they were erroneously included. Erroneous enumerations include duplicates, fictitious names, people who were born after census day or people who died prior to census day.

The authors are mathematical statisticians in the Decennial Statistical Studies Division of the U.S. Census Bureau. This paper reports the results of research and analysis undertaken by Census Bureau staff. It has undergone a more limited review than official Census Bureau Publications. This report is released to inform interested parties of research and to encourage discussion.

Table 1. P Sample and E Sample Comparison

	P Sample	E Sample
Estimates	Omissions	Erroneous Inclusions
Universe	All housing units in US[1]	Census housing units
PSUs	Block Clusters	Block Clusters

Block Cluster

The primary sampling units are block clusters, which are one or more geographically contiguous census blocks grouped together. Census blocks are formed by streets, roads, railroads, streams, etc. Forming block clusters involves a complicated hierarchical algorithm involving many rules and constraints. In general, the goal of block clustering is to produce sampling units that average about 30 housing units.

Integrated Coverage Measurement Survey

Until January 25, 1999, when the Supreme Court ruled that statistical sampling could not be used for the House of Representatives reapportionment, the Census Bureau had planned to conduct an Integrated Coverage Measurement (ICM) Survey. The primary goal of the ICM was to produce accurate and reliable direct state estimates, which would then be used for the reapportionment. Preliminary calculations indicated that the ICM allocation may result in coefficients of variation for the Dual System Estimate of approximately 0.5% in all states and standard errors of about 60,000 in the larger states (Schindler, 1998).

The Supreme Court ruling produced a change in the requirements. Direct state estimates were no longer needed for the reapportionment process, and consequently neither was a 750,000 housing unit sample. In contrast to the ICM, which incorporates the information into the population estimates, the A.C.E. results in a second set of estimates which will be used to evaluate the census and potentially for other purposes.

Because the Supreme Court ruling came too late to entirely redesign the sample, we will select an initial sample of block clusters using the ICM design. The independent list will be comprised of the housing units in these selected clusters, called the A.C.E. listing sample. The sample will be reduced during a later process called the A.C.E. Block Cluster Reduction. This has some limitations. The ICM was designed for efficient direct estimates for state total population. The primary goal for A.C.E., however, is to generate reliable demographic group estimates for the purpose of measuring differential coverage. The ICM sample is being selected using proportional allocation within a state. While this might be efficient for total population estimates, it is not efficient for estimating the population of smaller demographic groups. Overall, due to an increased sample size, we expect the reliability to be better for most of the poststrata estimates than the 1990 PES. Also, we expect the state total population estimates to be more reliable than for the 1990 PES.

Stratification and Sort Variables

Historically, coverage rates in the census have varied for many different groups in the population. In 1990, coverage rates were calculated for 357 poststrata identified by region, geographic area, race, Hispanic origin, age, sex, and tenure (own/rent). Although the estimated undercount for the total population was 1.6%, the estimated undercounts for the 357 groups ranged from -8.29% to 21.27% (Thompson, 1992). The poststrata definitions for Census 2000 are currently being researched and thus are not known. However, we are assuming they will be based on similar variables as in 1990 to account for the differential undercount. In order to estimate the coverage rates for several different poststrata with acceptable precision, there must be an adequate amount of sample selected for each of these poststrata. Since the characteristics of people within a block cluster vary, exact sample sizes for these groups are unattainable. However, the variation in the sample sizes for these groups can be improved by grouping similar block clusters together and selecting a systematic sample across these groups. In an attempt to better control the sample sizes from these different groups, block clusters will be classified into categories based on their estimated size, demographic composition, and level of urbanization.

Block clusters will initially be stratified into four mutually exclusive groups within each state: small block clusters (0-2 housing units), medium block clusters (3-79 housing units), large block clusters (80 or more housing units), and American Indian Reservation (AIR) block clusters. These groups will be sampled at different rates during the selection of the A.C.E. listing sample.

Although there will be no differential sampling within these four sampling strata, the clusters will be sorted by several variables in an attempt to sample a

[1]All housing units in the United States are eligible to be selected except housing units in Remote Alaska.

diverse set of block clusters. The first sort variable is the American Indian indicator, which has three categories:

- AIR or trustland
- tribal jurisdiction statistical area, Alaska Native Village statistical area or tribal designated statistical area
- all other areas

The second sort variable is the demographic group. Block clusters will be grouped with other block clusters containing similar demographic proportions based on 1990 census data. Assigning this variable to block clusters is described in more detail in the following paragraph. A third variable used for sorting the clusters is the level of urbanization. Each block cluster will be categorized as an urbanized area with 250,000 or more people, an urbanized area with less than 250,000 people, or a non-urban area. Finally, the clusters will be sorted geographically using county and cluster number.

To aid in selecting a sample that is well represented by the 6 major race/origin groups as well as owners and renters, block clusters will be classified into 12 demographic groups. Although many block clusters tend to have a large proportion of one demographic group, rarely are they entirely composed of only one, thus many clusters may fit well in two or more categories. To ensure that each cluster is assigned to only one group, a hierarchical assignment rule was developed so that when a cluster exceeds the group threshold, it is assigned to that group. These group thresholds were developed by grouping similar 1990 blocks together using a multivariate clustering method[2]. Table 2 lists these threshold values. The order of the hierarchy gives the smaller demographic groups priority over the larger ones and renters priority over owners.

A.C.E. Listing Sample Selection

For each state, a systematic sample is selected for each of the four strata listed in the previous section. In the following paragraphs, the sampling for the medium and large clusters is discussed, followed by the small block clusters and finally the AIR clusters.

As stated earlier, the Census Bureau was preparing to conduct an ICM during the early stages of the sample design. Thus the 25,000 block clusters were allocated to the states to approximately meet the ICM sample requirements, while maintaining a minimum of 300 block clusters per state. Selecting a sample of block clusters within each state results in approximately 2 million housing units to list. The sampling is done in two steps to guard against a listing workload that would be too formidable to complete in time. If the first systematic sample of block clusters results in a workload that is 10% more than the number of housing units allowed for listing, a second systematic sample is drawn from the first to approximately meet the listing constraint. Large block clusters are selected at a higher rate than medium clusters during the A.C.E. listing sample selection. These higher rates coupled with large block subsampling will result in more clusters represented in sample while keeping the total number of designated interviews within budget.

Table 2. Assignment Rule for Census 2000 A.C.E.

Order	Proportion	Threshold
1	Hawaiian and Pacific Islander Renters	0.10
2	Hawaiian and Pacific Islander Owners	0.10
3	American Indian and Alaska Native Renters	0.10
4	American Indian and Alaska Native Owners	0.10
5	Asian Renters	0.20
6	Asian Owners	0.20
7	Hispanic Renters	0.20
8	Hispanic Owners	0.20
9	Black Renters	0.25
10	Black Owners	0.25
11	White and other Renters	0.30
12	White and other Owners	all others

Small block clusters are generally sampled at a lower rate than both medium and large clusters. This is due to cost considerations which are further explained in a later section. These lower sampling rates cause some small cluster to have high weights, which may disproportionately affect the dual system estimates. In an attempt to avoid the problems associated with the high

[2]PROC FASTCLUS in SAS uses a multivariate clustering technique called nearest centroid sorting . For details, refer to pages 824-850 of the SAS/STAT User's Guide, Volume 1, Version 6, Fourth Edition.

weights we will initially sample 5,000 small block clusters. Using information about these 5,000 clusters we will attempt to target potential problem clusters in the subsampling operation which will reduce the number of small clusters in sample. These initial 5,000 small clusters were allocated to states proportionately to their projected number of housing units in small blocks. This allocation was bounded by two constraints -- a 20 block cluster minimum and a minimum expected sampling rate of 1 in 1000.

To ensure sufficient sample for calculating accurate undercount rates for American Indians on reservations, 355 block clusters will be selected from the block clusters on AIR nationwide. Small block clusters on AIR will not be included in this 355 block clusters. These clusters will be eligible for selection in the small cluster stratum. These 355 clusters were allocated to 26 states proportional to the 1990 population of American Indians on reservations. Ten states contained AIR clusters with little or no American Indian population. These clusters are not be included in an AIR stratum, but instead are eligible for selection in the other strata. The remaining 14 states and the District of Columbia contain no block clusters on AIR.

A.C.E. Block Cluster Reduction

As previously stated, the ICM sample will be reduced via the A.C.E. Block Cluster Reduction. This process is the first of three operations that will reduce the 2 million housing units listed down to approximately 300,000 housing units, which is nearly twice the sample size of the 1990 Post-Enumeration Survey (PES). The other two operations are described in the sections that follow. The sample was allocated to the states and the District of Columbia proportional to state population, with a minimum of 1,800 housing units designated for interview per state. The reduction will possibly have variable sampling rates within each state based on race, ethnicity and tenure classification of the block clusters. This differential sampling will help to provide sufficient sample sizes for providing estimates for several different poststrata. In order to provide sample for reliable AIR estimates, the AIR block clusters will not be reduced.

Small Block Cluster Subsampling

Small block clusters, those with between 0 and 2 housing units, get special attention in the A.C.E. These clusters have only a few housing units and are not a cost-effective workload for interviewing and follow-up operations. In order to wisely use our fixed resources we will sample small clusters at a lower rate than both medium and large clusters. Because of these uneven sampling rates the people in small clusters will have high weights. These high weights can disproportionately affect the dual system estimates. In 1990 only about 2.4% of the P sample people and 1.7% of the E sample people lived in small clusters. Yet these clusters contributed almost 10% to the net undercount and 15% to the estimated variance (Fay, 1998). In an attempt to improve our estimates we have developed a special design component to deal with small clusters.

Initially we will select 5,000 small clusters that will be a part of the A.C.E. address listing operation. Then through the small cluster subsampling operation we will reduce the number of small clusters in sample while at the same time attempting to achieve two other goals. First, we would like to prevent any small clusters from having weights that are extremely high compared to other clusters in the sample. Second we would like to limit the weights on the few clusters which we expected to be small, but turned out to be larger. Both of these goals would help to reduce the variance of the Dual System Estimator.

To achieve these goals we will use differential subsampling where the subsampling rates are based on the number of housing units on the Independent Listing and the number of housing units on the Census List. We are in the process of determining the methodology for attaining both goals.

Large Block Cluster Subsampling

Large block cluster subsampling is the final stage in selecting the housing units that are designated for an A.C.E. interview. The underlying concept of large block subsampling is to select a wide range of clusters, while still remaining within the budgeted number of housing units for interview. Assuming that people within a cluster are similar, interviewing all of them is not the most efficient use of resources. Instead, interviewing a smaller piece of several different clusters should provide a more geographically diverse sample.

This stage involves selecting a portion of each block cluster containing 80 or more housing units[3]. Housing units are selected by dividing each large cluster into segments of adjacent housing units, that differ by no more than one housing unit. Then, a sample of segments is selected by taking one systematic sample across all large clusters in a state. All housing units in the selected

---

[3]Clusters on American Indian Reservation are not subject to Large Block Cluster Subsampling.

segments are designated for A.C.E. interview. The sampling rate is determined so that the number of units selected for interview in large clusters added to the number selected in non-large clusters is approximately equal to the interviewing budget. In other words, since all housing units in non-large clusters are designated for interview, the difference between this number and the budgeted number of interviews is the target number of designated interviews from the large clusters.

E Sample Identification

Once the housing units have been selected for A.C.E. interview the next operation is to select the housing units that are in the E Sample. The information gathered from these housing units will be used to estimate the number of erroneous inclusions in the census. Although an overlapping P Sample and E Sample is not necessary, it is more cost efficient. If the E Sample includes many of the same people we can use the information from the P-sample interview to determine whether they were correctly enumerated and thus do not require a follow-up visit.

In an attempt to create overlapping samples, and thus save money, we will map the block clusters and segments of block clusters that are used to select the P Sample onto the census address list. If this step yields any cluster which will require more than 80 follow-up interviews, the E-sample housing units in these clusters will be subsampled.

Changes from Census 2000 Dress Rehearsal

In 1998, the Census Bureau conducted a Dress Rehearsal to refine the Census 2000 operations. The Dress Rehearsal revealed a few areas in the sample design that needed improvement. Many of the changes were minor operational details, but there are a few enhancements worth noting, two of which involve the treatment of small blocks.

The first change involves the formation of block clusters. Small blocks were not clustered with their neighbors for the Dress Rehearsal. Under certain conditions in 2000, small blocks are clustered with their neighbors. This reduces the total number of small clusters and thus reduces their weights. Overall, this change reduced the number of small clusters by about 65%, from 2,968,956 to 1,029,185. Under the new clustering procedure the initial weights for housing units in small clusters vary from 25 to 632 with an average of 221. Had improvements not been made, they would have ranged from 56 to 1,010 with an average of 588. Figure 1 shows the weight distributions of the 50 states, the District of Columbia, and Puerto Rico using both methods.

Also different in the Dress Rehearsal is the allocation of small clusters to states. In the Dress Rehearsal small clusters were allocated proportionately to the number of medium and large sample clusters in each site. This methodology is inefficient since many states have a large population but very little of it is contributed by small blocks whereas other states have a higher percentage of their population in small blocks. To account for this, in 2000 the small clusters were allocated proportional to the number of housing units projected in small clusters. This generally benefits states with larger proportions of the population residing in small clusters. The two allocations are listed in Table 3 for the states with the five highest and five lowest proportions of the population residing in small blocks.

Much of the A.C.E. operational planning was based on 1990 census data. For instance, the estimated number of housing units for creating the Independent List for each state was estimated based on 1990 information. Since these numbers were then used for renting office space and hiring staff in different areas of the country, exceeding these numbers may poise workload problems. Thus, these estimates became the listing constraints. To help keep the listing close to the listing constraints, two adjustments were built into the design. The first involves an adjustment prior to selecting a sample which is based on expected values. If it appears the listing would be too much based on the preliminary sampling rate, then the sampling rate was decreased. The second adjustment comes in the form of a two step sample. If the clusters selected during the first step surpass the listing constraint, a second sample from the first sample is selected. Without these two procedures, the listing would have surpassed the constraints by over 7.5 percent.

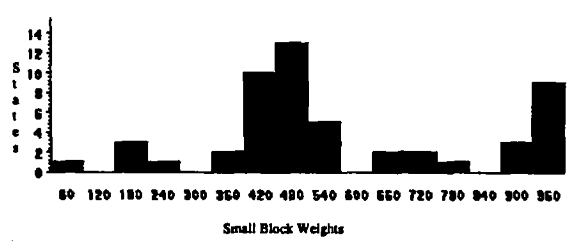
As can be seen by the sampling of changes listed in the above paragraphs, the A.C.E. sample design is continuously being updated and improved. Although there are still details to develop, such as the sampling rates for the small block subsampling and the possible strata for A.C.E. reduction, the framework is in place to provide reliable estimates of census coverage.

Table 3. Initial Small Block Cluster Weights for Selected States

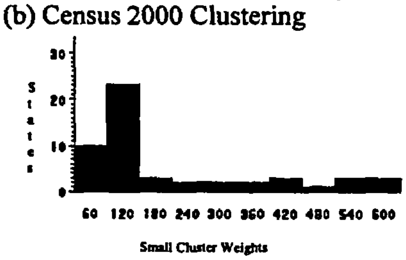
State	Percent 1990 Hus in Small Blocks	Dress Rehearsal Method Weight	Census 2000 Method Weight
North Dakota	11.67%	299	148
South Dakota	9.14%	246	139
Nebraska	5.47%	222	94
Kansas	4.64%	365	113
Wyoming	3.46%	529	617
Rhode Island	0.37%	11	41
New Jersey	0.32%	92	218
California	0.29%	156	467
Hawaii	0.24%	102	306
DC	0.06%	6	25

Figure 1. Frequency of Small Cluster Weights

(a) Dress Rehearsal Clustering



(b) Census 2000 Clustering



References

Bureau of the Census. (1997). Report to Congress: Plan for Census 2000. Washington, D.C.: Bureau of the Census.

Department of Commerce v. United States House of Representatives, No. 98-404 (U.S. filed Jan. 25, 1999).

Fay, R.E. (1988), "Evaluation of Census Coverage from the 1980 Post Enumeration Program (PEP): Census Omissions as Measured by the P Sample", Census Bureau Memorandum, March 10, 1988.

Fay, R. E. (1998), "Small Blocks in the 1990 PES", Census Bureau Memorandum, August 1998 (DRAFT).

SAS Institute Inc., *SAS/STAT User's Guide, Version 6, Fourth Edition, Volume 1*, Cary, NC: SS Institute Inc., 1989. 943 pp.

Schindler, E. (1998), "Allocation of the ICM Sample to the States for Census 2000," *Proceedings of Survey Research Methods Section, American Statistical Association*, Alexandria, VA, American Statistical Association, to appear.

Thompson, J. (1992), "CAPE Processing Results", Census Bureau Memorandum, March 20, 1992.